

Foundations of Applied Mathematics

Volume 1

Mathematical Analysis

JEFFREY HUMPHERYS

TYLER J. JARVIS

EMILY J. EVANS

BRIGHAM YOUNG UNIVERSITY



SOCIETY FOR INDUSTRIAL
AND APPLIED MATHEMATICS

PHILADELPHIA

Copyright © 2017 by the Society for Industrial and Applied Mathematics

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA.

No warranties, express or implied, are made by the publisher, authors, and their employers that the programs contained in this volume are free of error. They should not be relied on as the sole basis to solve a problem whose incorrect solution could result in injury to person or property. If the programs are employed in such a manner, it is at the user's own risk and the publisher, authors, and their employers disclaim all liability for such misuse.

Python is a registered trademark of Python Software Foundation.

PUBLISHER	David Marshall
EXECUTIVE EDITOR	Elizabeth Greenspan
DEVELOPMENTAL EDITOR	Gina Rinelli Harris
MANAGING EDITOR	Kelly Thomas
PRODUCTION EDITOR	Louis R. Primus
COPY EDITOR	Louis R. Primus
PRODUCTION MANAGER	Donna Witzleben
PRODUCTION COORDINATOR	Cally Shrader
COMPOSITOR	Lumina Datamatics
GRAPHIC DESIGNER	Lois Sellers
COVER DESIGNER	Sarah Kay Miller

Library of Congress Cataloging-in-Publication Data

Names: Humpherys, Jeffrey, author. | Jarvis, Tyler Jamison, author. | Evans, Emily J., author.

Title: Foundations of applied mathematics / Jeffrey Humpherys, Tyler J. Jarvis, Emily J. Evans, Brigham Young University, Provo, Utah.

Description: Philadelphia : Society for Industrial and Applied Mathematics, [2017]- | Series: Other titles in applied mathematics ; 152 | Includes bibliographical references and index.

Identifiers: LCCN 2017012783 | ISBN 9781611974898 (v. 1)

Subjects: LCSH: Calculus. | Mathematical analysis. | Matrices.

Classification: LOC QA303.2 .H86 2017 | DDC 515--dc23 LC record available at <https://lcn.loc.gov/2017012783>

siam is a registered trademark.

Contents

List of Notation	ix
Preface	xiii
I Linear Analysis I	1
1 Abstract Vector Spaces	3
1.1 Vector Algebra	3
1.2 Spans and Linear Independence	10
1.3 Products, Sums, and Complements	14
1.4 Dimension, Replacement, and Extension	17
1.5 Quotient Spaces	21
Exercises	27
2 Linear Transformations and Matrices	31
2.1 Basics of Linear Transformations I	32
2.2 Basics of Linear Transformations II	36
2.3 Rank, Nullity, and the First Isomorphism Theorem	40
2.4 Matrix Representations	46
2.5 Composition, Change of Basis, and Similarity	51
2.6 Important Example: Bernstein Polynomials	54
2.7 Linear Systems	58
2.8 Determinants I	65
2.9 Determinants II	70
Exercises	78
3 Inner Product Spaces	87
3.1 Introduction to Inner Products	88
3.2 Orthonormal Sets and Orthogonal Projections	94
3.3 Gram–Schmidt Orthonormalization	99
3.4 QR with Householder Transformations	105
3.5 Normed Linear Spaces	110
3.6 Important Norm Inequalities	117
3.7 Adjoints	120
3.8 Fundamental Subspaces of a Linear Transformation	123
3.9 Least Squares	127
Exercises	131

4	Spectral Theory	139
4.1	Eigenvalues and Eigenvectors	140
4.2	Invariant Subspaces	147
4.3	Diagonalization	150
4.4	Schur's Lemma	155
4.5	The Singular Value Decomposition	159
4.6	Consequences of the SVD	165
	Exercises	171
II	Nonlinear Analysis I	177
5	Metric Space Topology	179
5.1	Metric Spaces and Continuous Functions	180
5.2	Continuous Functions and Limits	185
5.3	Closed Sets, Sequences, and Convergence	190
5.4	Completeness and Uniform Continuity	195
5.5	Compactness	203
5.6	Uniform Convergence and Banach Spaces	210
5.7	The Continuous Linear Extension Theorem	213
5.8	Topologically Equivalent Metrics	219
5.9	Topological Properties	222
5.10	Banach-Valued Integration	227
	Exercises	233
6	Differentiation	241
6.1	The Directional Derivative	241
6.2	The Fréchet Derivative in $\mathbb{R}^n$	246
6.3	The General Fréchet Derivative	252
6.4	Properties of Derivatives	256
6.5	Mean Value Theorem and Fundamental Theorem of Calculus	260
6.6	Taylor's Theorem	265
	Exercises	272
7	Contraction Mappings and Applications	277
7.1	Contraction Mapping Principle	278
7.2	Uniform Contraction Mapping Principle	281
7.3	Newton's Method	285
7.4	The Implicit and Inverse Function Theorems	293
7.5	Conditioning	301
	Exercises	310
III	Nonlinear Analysis II	317
8	Integration I	319
8.1	Multivariable Integration	320
8.2	Overview of Daniell–Lebesgue Integration	326
8.3	Measure Zero and Measurability	331

8.4	Monotone Convergence and Integration on Unbounded Domains	335
8.5	Fatou's Lemma and the Dominated Convergence Theorem . . .	340
8.6	Fubini's Theorem and Leibniz's Integral Rule	344
8.7	Change of Variables	349
	Exercises	356
9	* Integration II	361
9.1	Every Normed Space Has a Unique Completion	361
9.2	More about Measure Zero	364
9.3	Lebesgue-Integrable Functions	367
9.4	Proof of Fubini's Theorem	372
9.5	Proof of the Change of Variables Theorem	374
	Exercises	378
10	Calculus on Manifolds	381
10.1	Curves and Arclength	381
10.2	Line Integrals	386
10.3	Parametrized Manifolds	389
10.4	* Integration on Manifolds	393
10.5	Green's Theorem	396
	Exercises	403
11	Complex Analysis	407
11.1	Holomorphic Functions	407
11.2	Properties and Examples	411
11.3	Contour Integrals	416
11.4	Cauchy's Integral Formula	424
11.5	Consequences of Cauchy's Integral Formula	429
11.6	Power Series and Laurent Series	433
11.7	The Residue Theorem	438
11.8	* The Argument Principle and Its Consequences	445
	Exercises	451
IV	Linear Analysis II	457
12	Spectral Calculus	459
12.1	Projections	460
12.2	Generalized Eigenvectors	465
12.3	The Resolvent	470
12.4	Spectral Resolution	475
12.5	Spectral Decomposition I	480
12.6	Spectral Decomposition II	483
12.7	Spectral Mapping Theorem	489
12.8	The Perron-Frobenius Theorem	494
12.9	The Drazin Inverse	500
12.10	* Jordan Canonical Form	506
	Exercises	511

13	Iterative Methods	519
13.1	Methods for Linear Systems	520
13.2	Minimal Polynomials and Krylov Subspaces	526
13.3	The Arnoldi Iteration and GMRES Methods	530
13.4	* Computing Eigenvalues I	538
13.5	* Computing Eigenvalues II	543
	Exercises	548
14	Spectra and Pseudospectra	553
14.1	The Pseudospectrum	554
14.2	Asymptotic and Transient Behavior	561
14.3	* Proof of the Kreiss Matrix Theorem	566
	Exercises	570
15	Rings and Polynomials	573
15.1	Definition and Examples	574
15.2	Euclidean Domains	583
15.3	The Fundamental Theorem of Arithmetic	588
15.4	Homomorphisms	592
15.5	Quotients and the First Isomorphism Theorem	598
15.6	The Chinese Remainder Theorem	601
15.7	Polynomial Interpolation and Spectral Decomposition	610
	Exercises	618
V	Appendices	625
A	Foundations of Abstract Mathematics	627
A.1	Sets and Relations	627
A.2	Functions	635
A.3	Orderings	643
A.4	Zorn's Lemma, the Axiom of Choice, and Well Ordering	647
A.5	Cardinality	648
B	The Complex Numbers and Other Fields	653
B.1	Complex Numbers	653
B.2	Fields	659
C	Topics in Matrix Analysis	663
C.1	Matrix Algebra	663
C.2	Block Matrices	665
C.3	Cross Products	667
D	The Greek Alphabet	669
	Bibliography	671
	Index	679

List of Notation

$\dagger$	indicates harder exercises xvi
$*$	indicates advanced material that can be skipped xvi
$\cong$	isomorphism 39, 595
$\oplus$	direct sum 16
$\boxplus$	addition in a quotient 24, 575, 599, 641
$\boxtimes$	multiplication in a quotient 24, 575, 599, 641
> 0	(for matrices) positive definite 159
≥ 0	(for matrices) positive semidefinite 159
$\preceq$	componentwise inequality 494, 513
$\langle \cdot, \cdot \rangle$	inner product 89
$(\cdot)$	the ideal generated by $\cdot$ 581
$\perp$	orthogonal complement 123
$\times$	Cartesian product 4, 14, 596, 630
$ $	divides 585
$ \cdot $	absolute value or modulus of a number; componentwise modulus of a matrix 4, 513, 654
$\ \cdot\ $	a norm 111
$\ \cdot\ _F$	the Frobenius norm 113, 115
$\ \cdot\ _{L^1}$	the L^1 -norm 134, 327
$\ \cdot\ _{L^2}$	the L^2 -norm 134
$\ \cdot\ _{L^\infty}$	the L^∞ -norm, also known as the sup norm 113, 134
$\ \cdot\ _{V,W}$	the induced norm on $\mathcal{B}(V,W)$ 114
$\ \cdot\ _p$	the p -norm, with $p \in [1, \infty]$, either for vectors or operators 111, 112, 115, 116
$[\![\cdot]\!]$	an equivalence class 22
$\llbracket \cdot, \cdot, \cdot \rrbracket$	a permutation 66
$\mathbb{1}$	the all ones vector $\mathbb{1} = [1 \ \dots \ 1]^T$ 499
$\mathbb{1}_E$	the indicator function of E 228, 323
2^S	the power set of S 628
$[a, b]$	a closed interval in $\mathbb{R}^n$ 321
a.e.	almost everywhere 329, 332
$\operatorname{argmin}_x f(x)$	the value of x that minimizes f 533

$B(\mathbf{x}_0, r)$	the open ball with center at $\mathbf{x}_0$ and radius r 182
B_j^n	the j th Bernstein polynomial of degree n 55
∂E	the boundary of E 192
$\mathcal{B}(V; W)$	the space of bounded linear transformations from V to W 114
$\mathcal{B}^k(V; W)$	the space of bounded linear transformations from V to $\mathcal{B}^{k-1}(V, W)$ 266
$\mathcal{B}(V)$	the space of bounded linear operators on V 114
$\mathcal{B}(X; \mathbb{F})$	the space of bounded linear functionals on X 214
$\mathbb{C}$	the complex numbers 4, 407, 627
$C(X; Y)$	the set of continuous functions from X to Y 5, 185
$C_0([a, b]; \mathbb{F})$	the space of continuous $\mathbb{F}$ -valued functions that vanish at the endpoints a and b 9
$C_b(X; \mathbb{R})$	the space of continuous functions on X with bounded L^∞ -norm 311
$C^n(X; Y)$	the space of Y -valued functions whose n th derivative is continuous on X 9, 253, 266
$C^\infty(X; Y)$	the space of smooth Y -valued functions on X 266
C_{ST}	the transition matrix from T to S 48
$D(\mathbf{x}_0, r)$	the closed ball with center at $\mathbf{x}_0$ and radius r 191
$Df(\mathbf{x})$	the Fréchet derivative of f at $\mathbf{x}$ 246
$D_i f$	the i th partial derivative of f 245
$D^k f(\mathbf{x})$	the k th Fréchet derivative of f at $\mathbf{x}$ 266
$D_{\mathbf{v}} f(\mathbf{x})$	the directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{v}$ 244
$D_{\mathbf{v}}^k f(\mathbf{x})$	k th directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{v}$ 268
$D^k f(\mathbf{x})\mathbf{v}^{(k)}$	same as $D_{\mathbf{v}}^k f(\mathbf{x})$ 268
D_λ	the eigennilpotent associated to the eigenvalue λ 483
$\text{diag}(\lambda_1, \dots, \lambda_n)$	the diagonal matrix with (i, i) entry equal to λ_i 152
$d(\mathbf{x}, \mathbf{y})$	a metric 180
δ_{ij}	the Kronecker delta 95
E°	the set of interior points of E 182
$\overline{E}$	the closure of E 192
$\mathcal{E}_\lambda$	the generalized eigenspace corresponding to eigenvalue λ 468
$\mathbf{e}_i$	the i th standard basis vector 13, 51
e_p	the evaluation map at p 33, 594
$\mathbb{F}$	a field, always either $\mathbb{C}$ or $\mathbb{R}$ 4
$\mathbb{F}^n$	n -dimensional Euclidean space 5
$\mathbb{F}[A]$	the ring of matrices that are polynomials in A 576

$\mathbb{F}[x]$	the space of polynomials with coefficients in $\mathbb{F}$ 6
$\mathbb{F}[x, y]$	the space of polynomials in x and y with coefficients in $\mathbb{F}$ 576
$\mathbb{F}[x; n]$	the space of polynomials with coefficients in $\mathbb{F}$ of degree at most n 9
$f^{-1}(U)$	the preimage $\{\mathbf{x} \mid f(\mathbf{x}) \in U\}$ of f 186, 187
$\gcd$	the greatest common divisor 586
Γ_f	the graph of f 635
$H_{\mathbf{v}}(\mathbf{x})$	the Householder transformation of $\mathbf{x}$ 106
$I(\gamma, z_0)$	the winding number of γ with respect to z_0 441
$\text{ind}(B)$	the index of the matrix B 466
$\Im(z)$	the imaginary part of z 450, 654
$K(A)$	the Kreiss matrix constant of A 562
$\mathcal{K}_k(A, \mathbf{b})$	the k th Krylov subspace of A generated by $\mathbf{b}$ 506, 527
$\kappa(A)$	the matrix condition number of A 307
κ	the relative condition number 303
$\hat{\kappa}$	the absolute condition number 302
$\kappa_{\text{spect}}(A)$	the spectral condition number of A 560
$L^1(A; X)$	the space of integrable functions on A 329, 338, 363
$L^p([a, b]; \mathbb{F})$	the space of p -integrable functions 6
L_i	the i th Lagrange interpolant 603
$L^\infty(S; X)$	the set of bounded functions from S to X , with the sup norm 216
$\mathcal{L}(V; W)$	the set of linear transformations from V to W 37
$\text{len}(\sigma)$	the arclength of the curve σ 383
$\ell(\mathbf{a}, \mathbf{b})$	the line segment from $\mathbf{a}$ to $\mathbf{b}$ 260
ℓ^p	the space of infinite sequences $(x_j)_{j=1}^\infty$ such that $\sum_{j=1}^\infty x_j^p$ converges 6
$\lambda(R)$	the measure of an interval $R \subset \mathbb{F}^n$ 323
$M_n(\mathbb{F})$	the space of square $n \times n$ matrices with entries in $\mathbb{F}$ 5
$M_{m \times n}(\mathbb{F})$	the space of $m \times n$ matrices with entries in $\mathbb{F}$ 5
$\mathbb{N}$	the natural numbers $\{0, 1, 2, \dots\}$ 577, 628
N_i	the i th Newton interpolant 611
$\mathbf{N}$	the unit normal 392
$\mathcal{N}(L)$	the kernel of L 35, 594
P_*	the eigenprojection associated to all the nonzero eigenvalues 500
P_λ	the eigenprojection associated to the eigenvalue λ 475
$p_A(z)$	the characteristic polynomial of A 143
$\text{proj}_X(\mathbf{v})$	the orthogonal projection of $\mathbf{v}$ onto the subspace X 96

$\text{proj}_{\mathbf{x}}(\mathbf{v})$	the orthogonal projection of $\mathbf{v}$ onto $\text{span}(\{\mathbf{x}\})$ 93
$\mathcal{P}$	a subdivision 323
π_i	the i th projection map from X^n to X 33
$\mathbb{Q}$	the rational numbers 628
Q_k	the points of $\mathbb{R}^n$ with dyadic rational coordinates 374
$\mathbb{R}$	the real numbers 4, 627
$R[[x]]$	the ring of formal power series in x with coefficients in R 576
$R(A, z)$	the resolvent of A , sometimes denoted $R(z)$ 470
$\text{Res}(f, z_0)$	the residue of f at z_0 440
$r(A)$	the spectral radius of A 474
$r_\varepsilon(A)$	the ε -pseudospectral radius of A 561
$\mathcal{R}([a, b]; X)$	the space of regulated integrable functions 324
$\mathcal{R}(L)$	the range of L 35
$\Re(z)$	the real part of z 450, 654
$\rho(L)$	the resolvent set of L 141
$S([a, b]; X)$	the set of all step functions mapping $[a, b]$ into X 228, 323
$\text{sign}(z)$	the complex sign $z/ z $ of z 656
$\text{Skew}_n(\mathbb{F})$	the space of skew-symmetric $n \times n$ matrices 5, 29
$\text{Sym}_n(\mathbb{F})$	the space of symmetric $n \times n$ matrices 5, 29
$\Sigma_\lambda(L)$	the λ -eigenspace of L 141
$\sigma(L)$	the spectrum of L 141
$\sigma_\varepsilon(A)$	the pseudospectrum of A 554
$\mathbf{T}$	the unit tangent vector 382
$T_{\mathbf{p}}M$	the tangent space to M at $\mathbf{p}$ 390
Θ_k	the set of compact n -cubes with corners in Q_k 374
$v(a)$	a valuation function 583
$\mathbb{Z}$	the integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$ 628
$\mathbb{Z}^+$	the positive integers $\{1, 2, \dots\}$ 628
$\mathbb{Z}_n$	the set of equivalence classes in $\mathbb{Z}$ modulo n 575, 633, 660

Preface

Overview

Why Mathematical Analysis?

Mathematical analysis is the foundation upon which nearly every area of applied mathematics is built. It is the language and intellectual framework for studying optimization, probability theory, stochastic processes, statistics, machine learning, differential equations, and control theory. It is also essential for rigorously describing the theoretical concepts of many quantitative fields, including computer science, economics, physics, and several areas of engineering.

Beyond its importance in these disciplines, mathematical analysis is also fundamental in the design, analysis, and optimization of algorithms. In addition to allowing us to make objectively true statements about the performance, complexity, and accuracy of algorithms, mathematical analysis has inspired many of the key insights needed to create, understand, and contextualize the fastest and most important algorithms discovered to date.

In recent years, the size, speed, and scale of computing has had a profound impact on nearly every area of science and technology. As future discoveries and innovations become more algorithmic, and therefore more computational, there will be tremendous opportunities for those who understand mathematical analysis. Those who can peer beyond the jargon-filled barriers of various quantitative disciplines and abstract out their fundamental algorithmic concepts will be able to move fluidly across quantitative disciplines and innovate at their crossroads. In short, mathematical analysis gives solutions to quantitative problems, and the future is promising for those who master this material.

To the Instructor

About this Text

This text modernizes and integrates a semester of advanced linear algebra with a semester of multivariable real analysis to give a new and redesigned year-long curriculum in linear and nonlinear analysis. The mathematical prerequisites are

vector calculus, linear algebra, and a semester of undergraduate-level, single-variable real analysis.¹

The content in this volume could be reasonably described as upper-division undergraduate or first-year graduate-level mathematics. It can be taught as a stand-alone two-semester sequence or in parallel with the second volume, *Foundations of Applied Mathematics, Volume 2: Algorithms, Approximation, and Optimization*, as part of a larger curriculum in applied and computational mathematics.

There is also a supplementary computer lab manual, containing over 25 computer labs that support this text. This text focuses on the theory, while the labs cover application and computation. Although we recommend that the manual be used in a computer lab setting with a teaching assistant, the material can be used without instruction. The concepts are developed slowly and thoroughly, with numerous examples and figures as pedagogical breadcrumbs, so that students can learn this material on their own, verifying their progress along the way. The labs and other classroom resources are open content and are available at

<http://www.siam.org/books/ot152>

The intent of this text and the computer labs is to attract and retain students into the mathematical sciences by modernizing the curriculum and connecting theory to application in a way that makes the students want to understand the theory, rather than just tolerate it. In short, a major goal of this text is to entice them to hunger for more.

Topics and Focus

In addition to standard material one would normally expect from linear and nonlinear analysis, this text also includes several key concepts of modern applied mathematical analysis which are not typically taught in a traditional applied math curriculum (see the Detailed Description, below, for more information).

We focus on both rigor and relevance to give the students mathematical maturity and an understanding of the most important ideas in mathematical analysis.

Detailed Description

Chapters 1–3 We give a rigorous treatment of the basics of linear algebra over both $\mathbb{R}$ and $\mathbb{C}$, including abstract vector spaces, linear transformations, matrices, the LU decomposition, inner product spaces, the QR decomposition, and least squares. As much as possible, we try to frame things in a way that does not require vector spaces to be finite dimensional, and we give many infinite-dimensional examples.

Chapter 4 We treat the spectral theory of matrices, including the spectral theorem for normal matrices. We give special attention to the singular value decomposition and its applications.

¹Specifically, the reader should have had exposure to a rigorous treatment of continuity, convergence, differentiation, and Riemann–Darboux integration in one dimension, as covered, for example, in [Abb15].

Chapter 5 We present the basics of metric topology, including the ideas of completeness and compactness. We define and give many examples of Banach spaces. Throughout the rest of the text we formulate results in terms of Banach spaces, wherever possible. A highlight of this chapter is the continuous linear extension theorem (sometimes called the bounded linear transformation theorem), which we use to give a very slick construction of Riemann (or rather *regulated*) Banach-valued integration (single-variable in this chapter and multivariable in Chapter 8).

Chapters 6–7 We discuss calculus on Banach spaces, including Fréchet derivatives and Taylor’s theorem. We then present the uniform contraction mapping theorem, which we use to prove convergence results for Newton’s method and also to give nice proofs of the inverse and implicit function theorems.

Chapters 8–9 We use the same basic ideas to develop Lebesgue integration as we used in the development of the regulated integral in Chapter 5. This approach could be called the *Riesz* or *Daniell* approach. Instead of developing measure theory and creating integrals from simple functions, we define what it means for a set to have zero measure and create integrals from step functions. This is a very clean way to do integration, which has the additional benefit of reinforcing many of the functional-analytic ideas developed earlier in the text.

Chapters 10–11 We give an introduction to the fundamental tools of complex analysis, briefly covering first the main ideas of parametrized curves, surfaces, and manifolds as well as line integrals, and Green’s theorem, to provide a solid foundation for contour integration and Cauchy’s theorem.

Throughout both of these chapters, we express the main ideas and results in terms of Banach-valued functions whenever possible, so that we can use these powerful tools to study spectral theory of operators later in the book.

Chapters 12–13 One of the biggest innovations in the book is our treatment of spectral theory. We take the Dunford–Schwartz approach via resolvents. This approach is usually only developed from an advanced functional-analytic point of view, but we break it down to the level of an undergraduate math major, using the tools and ideas developed earlier in this text.

In this setting, we put a strong emphasis on eigenprojections, providing insights into the spectral resolution theorem. This allows for easy proofs of the spectral mapping theorem, the Perron–Frobenius theorem, the Cayley–Hamilton theorem, and convergence of the power method. This also allows for a nice presentation of the Drazin inverse and matrix perturbation theory. These ideas are used again in Volume 4 with dynamical systems, where we prove the stable and center manifold theorems using spectral projections and corresponding semigroup estimates.

Chapter 14 The pseudospectrum is a fundamental tool in modern linear algebra. We use the pseudospectrum to study sequences of the form $\|A^k\|$, their asymptotic and transient behavior, an understanding of which is important both for Markov chains and for the many iterative methods based on such sequences, such as successive overrelaxation.

Chapter 15 We conclude the book with a chapter on applied ring theory, focused on the algebraic structure of polynomials and matrices. A major focus of this chapter is the Chinese remainder theorem, which we use in many ways, including to prove results about partial fractions and Lagrange interpolation. The highlight of the chapter is Section 15.7.3, which describes a striking connection between Lagrange interpolation and the spectral decomposition of a matrix.

Teaching from the Text

In our courses we teach each section in a fifty-minute lecture. We require students read the section carefully before each class so that class time can be used to focus on the parts they find most confusing, rather than on just repeating to them the material already written in the book.

There are roughly five to seven exercises per section. We believe that students can realistically be expected to do all of the exercises in the text, but some are difficult and will require time, effort, and perhaps an occasional hint. Exercises that are unusually hard are marked with the symbol †. Some of the exercises are marked with * to indicate that they cover advanced material. Although these are valuable, they are not essential for understanding the rest of the text, so they may safely be skipped, if necessary.

Throughout this book the exercises, examples, and concepts are tightly integrated and build upon each other in a way that reinforces previous ideas and prepares students for upcoming ideas. We find this helps students better retain and understand the concepts learned, and helps achieve greater depth. Students are encouraged to do *all* of the exercises, as they reinforce new ideas and also revisit the core ideas taught earlier in the text.

Courses Taught from This Book

Full Sequence

At BYU we teach a year-long advanced undergraduate-level course from this book, proceeding straight through the book, skipping only the advanced sections and chapters (marked with *), and ending in Chapter 13. But this would also make a very good course at the beginning graduate level as well.

Graduate students who are well prepared could be further challenged either by covering the material more rapidly, so as to get to the very rewarding material at the end of the book, or by covering some or all of the advanced sections along the way.

Advanced Linear Algebra

Alternatively, Chapters 1–4 (linear analysis part I), Section 7.5 (conditioning), and Chapters 12–14 (linear analysis part II), as well as parts of Chapter 15, as time permits, make up a very good one-semester advanced linear algebra course for students who have already completed undergraduate-level courses in linear algebra, complex analysis, and multivariate real analysis.

Advanced Analysis

This book can also be used to teach a one-semester advanced analysis course for students who have already had a semester of basic undergraduate analysis (say, at the level of [Abb15]). One possible path through the book for this course would be to briefly review Chapter 1 (vector spaces), Sections 2.1–2.2 (basics of linear transformations), and Sections 3.1 and 3.5 (inner product spaces and norms), in order to set notation and to remind the students of necessary background from linear algebra, and then proceed through Part II (Chapters 5–7) and Part III (Chapters 8–11).

Figure 1 indicates the dependencies among the chapters.

Advanced Sections

A few problems, sections, and even chapters are marked with the symbol * to indicate that they cover more advanced topics. Although this material is valuable, it is not essential for understanding the rest of the text, so it may safely be skipped, if necessary.

Instructors New to the Material

We’ve taken a tactical approach that combines professional development for faculty with instruction for the students. Specifically, the class instruction is where the theory lies and the supporting media (labs, etc.) are provided so that faculty need not be computer experts nor familiar with the applications in order to run the course.

The professor can teach the theoretical material in the text and use teaching assistants, who may be better versed in the latest technology, to cover the applications and computation in the labs, where the “hands-on” part of the course takes place. In this way the professor can gradually become acquainted with the applications and technology over time, by working through the labs on his or her own time without the pressures of staying ahead of the students.

A more technologically experienced applied mathematician could flip the class if she wanted to, or change it in other ways. But we feel the current format is most versatile and allows instructors of all backgrounds to gracefully learn and adapt to the program. Over time, instructors will become familiar enough with the content that they can experiment with various pedagogical approaches and make the program theirs.

To the Student

Examples

Although some of the topics in this book may seem familiar to you, especially many of the linear algebra topics, we have taken a very different approach to these topics by integrating many different topics together in our presentation, so examples treated in a discussion of vector spaces will appear again in sections on nonlinear analysis and other places throughout the text. Also, notation introduced in the examples is often used again later in the text.

Because of this, we recommend that you *read all the examples* in each section, even if the definitions, theorems, and other results look familiar.

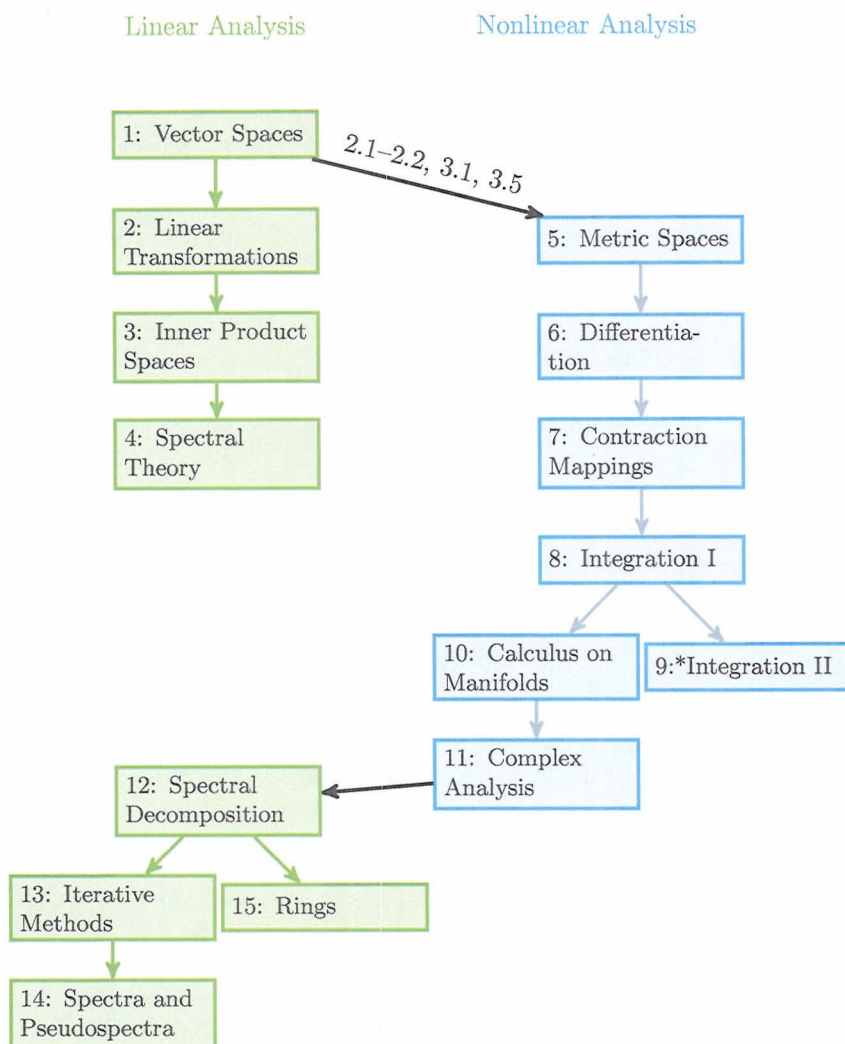


Figure 1. Diagram of the dependencies among the chapters of this book. Although we usually proceed straight through the book in order, it could also be used for either an advanced linear algebra course or a course in real and complex analysis. The linear analysis half (left side) provides a course in advanced linear algebra for students who have had complex analysis and multivariate real analysis. The nonlinear analysis half (right side) could be used for a course in real and complex analysis for students who have already had linear algebra and a semester of real analysis. For that track, we recommend briefly reviewing the material of Chapter 1 and Sections 2.1–2.2, 3.1, and 3.5, before proceeding to the nonlinear material, in order to fix notation and ensure students remember the necessary background.

Exercises

Each section of the book has several exercises, all collected at the end of each chapter. Horizontal lines separate the exercises for each section from the exercises for the other sections. We have carefully selected these exercises. You should work them all (but your instructor may choose to let you skip some of the advanced exercises marked with *)—each is important for your ability to understand subsequent material.

Although the exercises are gathered together at the end of the chapter, we strongly recommend that you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. Learning mathematics is like developing physical strength. It is much easier to improve, and improvement is greater, when exercises are done daily, in measured amounts, rather than doing long, intense bouts of exercise separated by long rests.

Origins

This curriculum evolved as an outgrowth of lecture notes and computer labs that were developed for a 6-credit summer course in computational mathematics and statistics. This was designed to introduce groups of undergraduate researchers to a number of core concepts in mathematics, statistics, and computation as part of a National Science Foundation (NSF) funded mentoring program called CSUMS: Computational Science Training for Undergraduates in the Mathematical Sciences.

This NSF program sought out new undergraduate mentoring models in the mathematical sciences, with particular attention paid to computational science training through genuine research experiences. Our answer was the Interdisciplinary Mentoring Program in Analysis, Computation, and Theory (IMPACT), which took cohorts of mathematics and statistics undergraduates and inserted them into an intense summer “boot camp” program designed to prepare them for interdisciplinary research during the school year. This effort required a great deal of experimentation, and when the dust finally settled, the list of topics that we wanted to teach blossomed into 8 semesters of material—essentially an entire curriculum.

After we explained the boot camp concept to one visitor, he quipped, “It’s the minimum number of instructions needed to create an applied mathematician.” Our goal, however, is much broader than this. We don’t want to train or create a specific type of applied mathematician; we want a curriculum that supports all types, simultaneously. In other words, our goal is to take in students with diverse and evolving interests and backgrounds and provide them with a common corpus of mathematical, statistical, and computational content so that they can emerge well prepared to work in *their own* chosen areas of specialization. We also want to draw their attention to the core ideas that are ubiquitous across various applications so that they can navigate fluidly across fields.

Acknowledgments

We thank the National Science Foundation for their support through the TUES Phase II grant DUE-1323785. We especially thank Ron Buckmire at the National

Science Foundation for taking a chance on us and providing much-needed advice and guidance along the way. Without the NSF, this book would not have been possible. We also thank the Department of Mathematics at Brigham Young University for their generous support and for providing a stimulating environment in which to work.

Many colleagues and friends have helped shape the ideas that led to this text, especially Randy Beard, Rick Evans, Shane Reese, Dennis Tolley, and Sean Warnick, as well as Bryant Angelos, Jonathan Baker, Blake Barker, Mylan Cook, Casey Dougal, Abe Frandsen, Ryan Grout, McKay Heasley, Amelia Henricksen, Ian Henricksen, Brent Kerby, Steven Lutz, Shane McQuarrie, Ryan Murray, Spencer Patty, Jared Webb, Matthew Webb, Jeremy West, and Alexander Zaitzeff, who were all instrumental in helping to organize this material.

We also thank the students of the BYU Applied and Computational Mathematics Emphasis (ACME) cohorts of 2013–2015, 2014–2016, 2015–2017, and 2016–2018, who suffered through our mistakes, corrected many errors, and never hesitated to tell us what they thought of our work.

We are deeply grateful to Chris Grant, Todd Kapitula, Zach Boyd, Rachel Webb, Jared McBride, and M. A. Averill, who read various drafts of this volume very carefully, corrected many errors, and gave us a tremendous amount of helpful feedback. Of course, all remaining errors are entirely our fault. We also thank Amelia Henricksen, Sierra Horst, and Michael Hansen for their help illustrating the text and Sarah Kay Miller for her outstanding graphic design work, including her beautifully designed book covers. We also appreciate the patience, support, and expert editorial work of Elizabeth Greenspan and the other editors and staff at SIAM.

Finally, we thank the folks at Savvysherpa, Inc., for corporate sponsorship that greatly helped make the transition from IMPACT to ACME and their help nourishing and strengthening the ACME development team.

Part I

Linear Analysis I

1

Abstract Vector Spaces

Mathematics is the art of reducing any problem to linear algebra.
—William Stein

In this chapter we develop the theory of abstract vector spaces. Although it is sometimes sufficient to think of vectors as arrows or arrays in $\mathbb{R}^n$, this point of view can be too limiting, particularly when it comes to non-Cartesian coordinates, infinite-dimensional vector spaces, or vectors over other fields, such as the complex numbers $\mathbb{C}$. Hence, we begin this journey by stripping away the simple geometric interpretation of a vector space and replacing it with rigorous mathematical abstraction, from which we build the theory of mathematical analysis. We then carefully rebuild the geometric properties of vector spaces in subsequent chapters.

This chapter contains a description of vector spaces, subspaces, and the rules of vector algebra. In addition, we discuss linear combinations of vectors, the spans of sets, and the consequences of linear independence. We then build new vector spaces and subspaces out of existing ones by taking Cartesian products of vector spaces and direct sums of subspaces. The chapter concludes with an important foray into quotient spaces and their algebraic properties.

Even if you are already familiar with linear algebra, we recommend that you at least read the examples and work the exercises, because they are used and referred to repeatedly throughout this text.

1.1 Vector Algebra

Vector spaces are a fundamental concept in mathematics, science, and engineering. The key properties of a vector space are that its elements (called vectors) can be added, subtracted, and rescaled.

We can describe many real-world phenomena as vectors in a vector space. Examples include the position and momentum of particles, sound traveling in a medium, and electrical or optical signals from an Internet connection. Once we describe something as a vector, the powerful tools of linear algebra can be used to study it.

Remark 1.1.1. Many of the properties of vector spaces described in this text hold over arbitrary fields;² however, we restrict our discussion here to vector spaces over the real field $\mathbb{R}$ or the complex³ field $\mathbb{C}$. We denote the field by $\mathbb{F}$ when a statement is true for both $\mathbb{R}$ and $\mathbb{C}$.

Remark 1.1.2. We use the notation $|\cdot|$ to denote the absolute value of a real number and the modulus of a complex number; that is, if $z = a + bi \in \mathbb{C}$, where $a, b \in \mathbb{R}$, then $|z| = \sqrt{a^2 + b^2}$. The reader should verify that $|z|^2 = z\bar{z} = \bar{z}z$, where $\bar{z} = \overline{a + bi} = a - bi$ is the complex conjugate.

1.1.1 Vector Spaces

We begin by carefully defining a vector space.

Definition 1.1.3. A vector space over a field $\mathbb{F}$ is a set V with two operations: addition, mapping the Cartesian product⁴ $V \times V$ to V , and denoted by $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x} + \mathbf{y}$; and scalar multiplication, mapping $\mathbb{F} \times V$ to V , and denoted by $(a, \mathbf{x}) \mapsto a\mathbf{x}$. Elements of the vector space V are called vectors, and the elements of the field $\mathbb{F}$ are called scalars. These operations must satisfy the following properties for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and all $a, b \in \mathbb{F}$:

- (i) *Commutativity of vector addition:* $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
- (ii) *Associativity of vector addition:* $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
- (iii) *Existence of an additive identity:* There exists an element $\mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x}$.
- (iv) *Existence of an additive inverse:* For each $\mathbf{x} \in V$ there exists an element $-\mathbf{x} \in V$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
- (v) *First distributive law:* $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$.
- (vi) *Second distributive law:* $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$.
- (vii) *Multiplicative identity:* $1\mathbf{x} = \mathbf{x}$.
- (viii) *Relation to ordinary multiplication:* $(ab)\mathbf{x} = a(b\mathbf{x})$.

Nota Bene 1.1.4. A subtle point that is sometimes missed because it is not included in the numbered list of properties is that the definition of a vector space requires the operations of vector addition and scalar multiplication to take their values in V . More precisely, if we add two vectors together in V , the result must be a vector in V , and if we multiply a vector in V by a scalar in

²For more information on general fields, see Appendix B.2.

³For more information on complex numbers, see Appendix B.1.

⁴See Definition A.1.10 (viii) in Appendix A.

$\mathbb{F}$, the result must also be a vector in V . When these properties hold, we say that V is *closed* under vector addition and scalar multiplication. If V is not closed under an operation, then, strictly speaking, we don't have an operation on V at all. Checking closure under operations is often the hardest part of verifying that a given set is a vector space.

Remark 1.1.5. Throughout this text we typically write vectors in boldface in an effort to distinguish them from scalars and other mathematical objects. This may cause confusion, however, when the vectors are also functions, polynomials, or other objects that are normally written in regular (not bold) fonts. In such cases, we simply write these vectors in the regular font and expect the reader to recognize the objects as vectors.

Example 1.1.6. Each of the following examples of vector spaces is important in applied mathematics and appears repeatedly throughout this book. The naïve descriptions of vectors as “arrows” with “magnitude” and “direction” are inadequate for describing many of these examples. The reader should check that all the axioms are satisfied for each of these:

- (i) The n -tuples $\mathbb{F}^n$ forming the usual Euclidean space. Vector addition is given by $(a_1, \dots, a_n) + (b_1, \dots, b_n) = (a_1 + b_1, \dots, a_n + b_n)$, and scalar multiplication is given by $c(a_1, \dots, a_n) = (ca_1, \dots, ca_n)$.
- (ii) The space of $m \times n$ matrices, denoted $M_{m \times n}(\mathbb{F})$, where each entry is an element of $\mathbb{F}$. Vector addition is usual matrix addition, and scalar multiplication is given by multiplying every entry in the matrix by the scalar:

$$c \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} ca_{11} & \dots & ca_{1n} \\ \vdots & \ddots & \vdots \\ ca_{m1} & \dots & ca_{mn} \end{pmatrix}.$$

We often write $M_n(\mathbb{F})$ to denote square $n \times n$ matrices. There is another important operation called *multiplication* on this set, namely, matrix multiplication. Matrix multiplication is important in many contexts, but it is not part of the vector space structure—that is, $M_n(\mathbb{F})$ and $M_{m \times n}(\mathbb{F})$ are vector spaces regardless of whether we can multiply matrices.

- (iii) For $[a, b] \subset \mathbb{R}$, the space $C([a, b]; \mathbb{F})$ of continuous $\mathbb{F}$ -valued functions. Vector addition is given by defining the function $f + g$ as $(f + g)(x) = f(x) + g(x)$, and scalar multiplication is given by defining the function cf by $(cf)(x) = c \cdot f(x)$. Note that $C([a, b]; \mathbb{F})$ is closed under vector addition and scalar multiplication because sums and scalar products of continuous functions are continuous.

(iv) For $[a, b] \subset \mathbb{R}$ and $1 \leq p < \infty$, the space $L^p([a, b]; \mathbb{F})$ of p -integrable^a functions $f : [a, b] \rightarrow \mathbb{F}$, that is, satisfying $\int_a^b |f(x)|^p dx < \infty$.

For $p = \infty$, let $L^\infty([a, b]; \mathbb{F})$ be the set of all functions $f : [a, b] \rightarrow \mathbb{F}$ such that $\sup_{x \in [a, b]} |f(x)| < \infty$.

If $1 \leq p \leq \infty$, the space $L^p([a, b]; \mathbb{F})$ is a vector space. To prove that L^p is closed under vector addition, notice that for $a, b \in \mathbb{R}$ we have

$$|a + b|^p \leq (2 \max\{|a|, |b|\})^p \leq 2^p(|a|^p + |b|^p).$$

Thus, it follows that

$$\int_a^b |f(x) + g(x)|^p dx \leq 2^p \left(\int_a^b |f(x)|^p dx + \int_a^b |g(x)|^p dx \right).$$

(v) The space $\mathbb{F}[x]$ of polynomials with coefficients in $\mathbb{F}$.

(vi) The space ℓ^p of infinite sequences $(x_j)_{j=1}^\infty$ with each $x_j \in \mathbb{F}$ and satisfying $\sum_{j=1}^\infty |x_j|^p < \infty$ for $1 \leq p < \infty$, and satisfying $\sup_{j \in \mathbb{N}} |x_j| < \infty$ for $p = \infty$. Vector addition and scalar multiplication are defined componentwise. It is not immediately obvious that ℓ^p is closed under vector addition, but this can be shown with an argument similar to the one used for $L^p([a, b]; \mathbb{F})$ in (iv). We see another proof when we discuss Minkowski's inequality in Section 3.6.

^aFor now we use the Riemann integral in the definition of L^p , but this is an abuse of notation, since the usual definition of L^p assumes a more advanced definition of integration. In Chapters 8–9 we generalize the integral to include a broader class of integrable functions consistent with the usual convention.

Proposition 1.1.7. *Let V be a vector space. If $\mathbf{x}, \mathbf{y} \in V$, then the following hold:*

- (i) $\mathbf{x} + \mathbf{y} = \mathbf{x}$ implies $\mathbf{y} = \mathbf{0}$. In particular, the additive identity is unique.
- (ii) $\mathbf{x} + \mathbf{y} = \mathbf{0}$ implies $\mathbf{y} = -\mathbf{x}$. In particular, additive inverses are unique.
- (iii) $0\mathbf{x} = \mathbf{0}$.
- (iv) $(-1)\mathbf{x} = -\mathbf{x}$.

Proof.

- (i) If $\mathbf{x} + \mathbf{y} = \mathbf{x}$, then $\mathbf{0} = -\mathbf{x} + \mathbf{x} = -\mathbf{x} + (\mathbf{x} + \mathbf{y}) = (-\mathbf{x} + \mathbf{x}) + \mathbf{y} = \mathbf{0} + \mathbf{y} = \mathbf{y}$.
- (ii) If $\mathbf{x} + \mathbf{y} = \mathbf{0}$, then $-\mathbf{x} = -\mathbf{x} + \mathbf{0} = -\mathbf{x} + (\mathbf{x} + \mathbf{y}) = (-\mathbf{x} + \mathbf{x}) + \mathbf{y} = \mathbf{0} + \mathbf{y} = \mathbf{y}$.
- (iii) For each $\mathbf{x}$, we have that $\mathbf{x} = 1\mathbf{x} = (1 + 0)\mathbf{x} = 1\mathbf{x} + 0\mathbf{x} = \mathbf{x} + 0\mathbf{x}$. Hence, (i) implies that $0\mathbf{x} = \mathbf{0}$.
- (iv) For each $\mathbf{x}$, we have that $\mathbf{0} = 0\mathbf{x} = (1 + (-1))\mathbf{x} = 1\mathbf{x} + (-1)\mathbf{x} = \mathbf{x} + (-1)\mathbf{x}$. Hence, (ii) implies that $(-1)\mathbf{x} = -\mathbf{x}$. $\square$

Remark 1.1.8. Subtraction in a vector space is really just addition of the negative. Thus, the expression $\mathbf{x} - \mathbf{y}$ is just shorthand for $\mathbf{x} + (-\mathbf{y})$.

Proposition 1.1.9. *Let V be a vector space. If $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $a, b \in \mathbb{F}$, then the following hold:*

- (i) $a\mathbf{0} = \mathbf{0}$.
- (ii) If $a\mathbf{x} = a\mathbf{y}$ and $a \neq 0$, then $\mathbf{x} = \mathbf{y}$.
- (iii) If $\mathbf{x} + \mathbf{y} = \mathbf{x} + \mathbf{z}$, then $\mathbf{y} = \mathbf{z}$.
- (iv) If $a\mathbf{x} = b\mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$, then $a = b$.

Proof.

- (i) For each $\mathbf{x} \in V$ and $a \in \mathbb{F}$, we have that $a\mathbf{x} = a(\mathbf{x} + \mathbf{0}) = a\mathbf{x} + a\mathbf{0}$. Since the additive identity is unique by Proposition 1.1.7, it follows that $a\mathbf{0} = \mathbf{0}$.
- (ii) Since $a \neq 0$, we have that $a^{-1} \in \mathbb{F}$. Thus, $\mathbf{x} = (1\mathbf{x}) = (a^{-1}a)\mathbf{x} = a^{-1}(a\mathbf{x}) = a^{-1}(a\mathbf{y}) = (a^{-1}a)\mathbf{y} = 1\mathbf{y} = \mathbf{y}$.
- (iii) Note that $\mathbf{y} = \mathbf{0} + \mathbf{y} = (-\mathbf{x} + \mathbf{x}) + \mathbf{y} = -\mathbf{x} + (\mathbf{x} + \mathbf{y}) = -\mathbf{x} + (\mathbf{x} + \mathbf{z}) = (-\mathbf{x} + \mathbf{x}) + \mathbf{z} = \mathbf{0} + \mathbf{z} = \mathbf{z}$.
- (iv) We prove an equivalent statement: If $c\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$, then $c = 0$. By way of contradiction, assume $c \neq 0$. It follows that $\mathbf{x} = 1\mathbf{x} = (c^{-1}c)\mathbf{x} = c^{-1}(c\mathbf{x}) = c^{-1}\mathbf{0} = \mathbf{0}$, which is a contradiction. Hence, $c = 0$. $\square$

Nota Bene 1.1.10. It is important to understand that Proposition 1.1.9(iv) is not saying that we can divide one vector by another vector. That is absolutely not the case. Indeed, this cancellation only works in the special case that both sides of the equation are scalar multiples of the same vector $\mathbf{x}$.

1.1.2 Subspaces

We conclude this section by defining subspaces of a vector space and then providing several examples.

Definition 1.1.11. *Let V be a vector space. A nonempty subset $W \subset V$ is a subspace of V if W itself is a vector space under the same operations of vector addition and scalar multiplication as V .*

An immediate consequence of the definition is that if a subset $W \subset V$ is a subspace, then it is closed under vector addition and scalar multiplication as defined by V . In other words, if W is a subspace, then whenever $\mathbf{x}, \mathbf{y} \in W$ and $a, b \in \mathbb{F}$, we have that $a\mathbf{x} + b\mathbf{y} \in W$. Surprisingly, this is all we need to check to show that a subset is a subspace.

Theorem 1.1.12. *If W is a nonempty subset of a vector space V such that for any $\mathbf{x}, \mathbf{y} \in W$ and for any $a, b \in \mathbb{F}$ the vector $a\mathbf{x} + b\mathbf{y}$ is also in W , then W is a subspace of V .*

Proof. The hypothesis of this theorem shows that W is closed under vector addition and scalar multiplication, so vector addition does indeed map $W \times W$ to W and scalar multiplication does indeed map $\mathbb{F} \times W$ to W , as required. Properties (i)–(ii) and (v)–(viii) of Definition 1.1.3 follow because they hold for the space V . The proofs of the remaining two properties (iii) and (iv) are left as an exercise—see Exercise 1.2. $\square$

The following corollary is immediate.

Corollary 1.1.13. *If V is a vector space and W is a nonempty subset of V such that for any $\mathbf{x}, \mathbf{y} \in W$ and for any $a \in \mathbb{F}$ the vectors $\mathbf{x} + \mathbf{y}$ and $a\mathbf{x}$ are also in W , then W is a subspace of V .*

Unexample 1.1.14.

- (i) If W is the union of the x - and y -axes in the plane $\mathbb{R}^2$, then it is *not* a subspace of $\mathbb{R}^2$. Indeed, the sum of the vectors $(1, 0)$ (in the x -axis) and $(0, 1)$ (in the y -axis) is $(1, 1)$, which is not in W , so W is not closed with respect to vector addition.
- (ii) Any line or plane in $\mathbb{R}^3$ that does not contain the origin $\mathbf{0}$ is not a subspace of $\mathbb{R}^3$ because all subspaces are required to contain $\mathbf{0}$.

Example 1.1.15. The following are examples of subspaces:

- (i) Any line that passes through the origin in $\mathbb{R}^3$ is a subspace of $\mathbb{R}^3$ and can be written as $\{t\mathbf{x} \mid t \in \mathbb{R}\}$ for some $\mathbf{x} \in \mathbb{R}^3$. Thus, any scalar multiple $a(t\mathbf{x})$ is on the line, as is any sum $t\mathbf{x} + s\mathbf{x} = (t + s)\mathbf{x}$. More generally, for any vector space V over a field $\mathbb{F}$ and any $\mathbf{x} \in V$, the set $W = \{t\mathbf{x} \mid t \in \mathbb{F}\}$ is a subspace of V .

- (ii) Any plane that passes through the origin in $\mathbb{R}^3$ is a subspace of $\mathbb{R}^3$ and can be written as the set $\{s\mathbf{x} + t\mathbf{y} \mid s, t \in \mathbb{R}\}$ for some pair^a of vectors $\mathbf{x}$ and $\mathbf{y}$. It is straightforward to check that this set is closed under vector addition and scalar multiplication. More generally, for any vector space V over a field $\mathbb{F}$ and any two vectors $\mathbf{x}, \mathbf{y} \in V$, the set $P = \{s\mathbf{x} + t\mathbf{y} \mid s, t \in \mathbb{F}\}$ is a subspace of V .
- (iii) For $[a, b] \subset \mathbb{R}$, the space $C_0([a, b]; \mathbb{F})$ of all functions $f \in C([a, b]; \mathbb{F})$ such that $f(a) = f(b) = 0$ is a subspace of $C([a, b]; \mathbb{F})$. To see this, we must check that if f and g both vanish at the endpoints, then so does $cf + dg$ for any $c, d \in \mathbb{F}$.
- (iv) For $n \in \mathbb{N}$, the set $\mathbb{F}[x; n]$ of all polynomials of degree at most n is a subspace of $\mathbb{F}[x]$. To see this, we must check that for any f and g of degree at most n and any $a, b \in \mathbb{F}$, then $af + bg$ is again a polynomial of degree at most n .
- (v) For any vector space V , both $\{\mathbf{0}\}$ and V are subspaces of V . The subspace $\{\mathbf{0}\}$ is called the *trivial* subspace of V . Any subspace of V that is not equal to V itself is called a *proper* subspace.
- (vi) For $[a, b] \subset \mathbb{R}$, the set $C([a, b]; \mathbb{F})$ is a subspace of $L^p([a, b]; \mathbb{F})$ when $1 \leq p < \infty$.
- (vii) The space $C^n((a, b); \mathbb{F})$ of $\mathbb{F}$ -valued functions whose n th derivative is continuous on (a, b) is a subspace of $C((a, b); \mathbb{F})$.
- (viii) The property of being a subspace is transitive. If X is a subspace of a vector space W , and if W is a subspace of a vector space V , then X is a subspace of V . For example, when $n \in \mathbb{N}$, the space $C^{n+1}((a, b); \mathbb{F})$ is a subspace of $C^n((a, b); \mathbb{F})$. It follows that $C^{n+1}((a, b); \mathbb{F})$ is a subspace of $C((a, b); \mathbb{F})$.

^aIf the two vectors are scalar multiples of each other, then this set is a line, not a plane.

Application 1.1.16 (Linear Maps and the Superposition Principle).

Many physical phenomena satisfy the *superposition principle*, which is just another way of saying that they are vectors in a vector space. Waves—both electromagnetic and acoustic—provide common examples of this phenomenon. Radio signals are examples of electromagnetic waves. Two radio signals of different frequencies correspond to two different vectors in a common vector space. When these are broadcast simultaneously, the result is a new wave (also a vector) that is the sum of the two original signals.

Sound waves also behave like vectors in a vector space. This allows the construction of technologies like noise-canceling headphones. The idea is very simple. When an undesired noise $\mathbf{n}$ is approaching the ear, produce the signal that is the additive inverse $-\mathbf{n}$, and play it at the same time. The signal heard is the sum $\mathbf{n} + (-\mathbf{n}) = \mathbf{0}$, which is silent.

1.2 Spans and Linear Independence

In this section we discuss two big ideas from linear algebra. The first is the *span* of a collection of vectors in a given vector space V . The span of a nonempty set S is the set of all *linear combinations* of elements from S , that is, the set of all vectors obtained by taking finite sums of scalar multiples of the elements of S . This set of linear combinations forms a subspace of V , which can also be obtained by taking the intersection of all subspaces of V that contain S .

The second idea concerns whether it is possible to remove a vector from S and still span the same subspace. If no vectors can be removed, then the set S is *linearly independent*, and any vector in the span can be uniquely represented as a linear combination of vectors from S .

If S is linearly independent and the span of S is the entire vector space V , then every vector in V has a unique representation as a linear combination of elements from S . In this case S defines a coordinate system for the vector space, and we say that S is a *basis* of V . Such a representation is extremely useful and forms a very powerful set of tools with incredibly broad application.

Throughout the rest of this chapter, we assume that S is a subset of a given vector space V , unless otherwise specified.

1.2.1 The Span of a Set

The span of a collection of vectors is an important building block in linear algebra. The main result of this subsection is that the span of a set can be defined two ways: either as the intersection of all subspaces that contain the set, or as the set of all possible linear combinations of the set.

Definition 1.2.1. *The span of S , denoted $\text{span}(S)$, is the set of linear combinations of elements of S , that is, the set of all finite sums of the form*

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m, \quad \mathbf{x}_i \in S, \quad a_i \in \mathbb{F}, \quad m \in \mathbb{N}. \quad (1.1)$$

If S is empty, then we define $\text{span}(S)$ to be the set $\{\mathbf{0}\}$. If $\text{span}(S) = W$ for some subspace W of V , then we say that S spans W .

Proposition 1.2.2. *The set $\text{span}(S)$ is a subspace of V .*

Proof. If S is empty, the result is clear, so we may assume S is nonempty. If $\mathbf{x}, \mathbf{y} \in \text{span}(S)$, then there exists a finite subset $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset S$, such that $\mathbf{x} = \sum_{i=1}^m c_i \mathbf{x}_i$ and $\mathbf{y} = \sum_{i=1}^m d_i \mathbf{x}_i$ for some coefficients $c_1, \dots, c_m$ and $d_1, \dots, d_m$ (some

possibly zero). Since $ax + by = \sum_{i=1}^m (ac_i + bd_i)x_i$ is of the form given in (1.1), and is thus a linear combination of elements of S , it follows that $ax + by$ is contained in S ; hence $\text{span}(S)$ is a subspace of V . $\square$

Lemma 1.2.3. *If W is a subspace of V , then $\text{span}(W) = W$.*

Proof. It is immediate that $W \subset \text{span}(W)$, so it suffices to show $\text{span}(W) \subset W$, which we prove by induction. By definition, any element $\mathbf{v} \in \text{span}(W)$ can be written as a linear combination $\mathbf{v} = a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n$ with $\mathbf{x}_i \in W$. If $n = 1$, then $\mathbf{v} \in W$ since subspaces are closed under scalar multiplication. Now suppose that all linear combinations of W of length k or less are in W , and consider a linear combination of length $k + 1$, say $\mathbf{v} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k + a_{k+1}\mathbf{x}_{k+1}$. This can be rewritten as $\mathbf{v} = \mathbf{w} + a_{k+1}\mathbf{x}_{k+1}$, where $\mathbf{w}$ is the sum of the first k terms in the linear combination. By the inductive hypothesis, $\mathbf{w} \in W$, and by the definition of a subspace, $\mathbf{w} + a_{k+1}\mathbf{x}_{k+1} \in W$. Thus, all linear combinations of $k + 1$ elements are also in W . Therefore, by induction, all finite linear combinations of elements of W are in W , or in other words $\text{span}(W) \subset W$. $\square$

Proposition 1.2.4. *The intersection of a collection $\{W_\alpha\}_{\alpha \in J}$ of subspaces of V is a subspace of V .*

Proof. The intersection is nonempty because it contains $\mathbf{0}$. Assume W_α is a subspace of V for each $\alpha \in J$. If $\mathbf{x}, \mathbf{y} \in \bigcap_{\alpha \in J} W_\alpha$, then $\mathbf{x}, \mathbf{y} \in W_\alpha$ for each $\alpha \in J$. Since each W_α is a subspace, then for every $a, b \in \mathbb{F}$ we have $a\mathbf{x} + b\mathbf{y} \in W_\alpha$. Hence, $a\mathbf{x} + b\mathbf{y} \in \bigcap_{\alpha \in J} W_\alpha$, and, by Theorem 1.1.12, we conclude that $\bigcap_{\alpha \in J} W_\alpha$ is a subspace of V . $\square$

Proposition 1.2.5. *If $S \subset S' \subset V$, then $\text{span}(S) \subset \text{span}(S')$.*

Proof. See Exercise 1.10. $\square$

Theorem 1.2.6. *The span of a set $S \subset V$ is the smallest subspace of V that contains S , meaning that for any subspace $W \subset V$ containing S we have $\text{span}(S) \subset W$. It is also equal to the intersection of all subspaces that contain S .*

Proof. If W is any subspace of V with $S \subset W$, then $\text{span}(S) \subset \text{span}(W) = W$ by Lemma 1.2.3. Moreover, the intersection of all subspaces containing S is a subspace containing S , by Proposition 1.2.4, and hence it contains $\text{span}(S)$. Conversely, $\text{span}(S)$ is a subspace of V containing S , so it must contain the intersection of all such subspaces. $\square$

Proposition 1.2.7. *If $\mathbf{v} \in \text{span}(S)$, then $\text{span}(S) = \text{span}(S \cup \{\mathbf{v}\})$.*

Proof. By the previous proposition, we have that $\text{span}(S) \subset \text{span}(S \cup \{\mathbf{v}\})$. Thus, assuming the hypothesis, it suffices to show that $\text{span}(S \cup \{\mathbf{v}\}) \subset \text{span}(S)$.

Given $\mathbf{x} \in \text{span}(S \cup \{\mathbf{v}\})$ we have $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{s}_i + c\mathbf{v}$, for some $\{\mathbf{s}_i\}_{i=1}^n \subset S$. Since $\mathbf{v} \in \text{span}(S)$, we also have that $\mathbf{v} = \sum_{j=1}^m b_j \mathbf{t}_j$ for some $\{\mathbf{t}_j\}_{j=1}^m \subset S$. Thus, we have that $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{s}_i + c \sum_{j=1}^m b_j \mathbf{t}_j$, which is a linear combination of elements of S . Therefore, $\mathbf{x} \in \text{span}(S)$. $\square$

1.2.2 Linear Independence

As mentioned in the introduction to this section, a set S is linearly independent if it is impossible to remove a vector from S and still span the same subspace.

The first result in this subsection establishes that any vector in the span of a linearly independent set S can be uniquely expressed as a finite sum of scalar multiples of vectors from S . We then prove that subsets of linearly independent sets are also linearly independent. One might say that this result is dual to Proposition 1.2.5, which states that supersets of sets that span also span. These two dual results intersect when S both spans V and is linearly independent—if we add a vector to S , then it still spans V but is no longer linearly independent; if we remove a vector from S , then it is still linearly independent but no longer spans V . When S is both linearly independent and spans the space, we say that it is a *basis* for V .

Definition 1.2.8. *The set S is linearly dependent if there exists a nontrivial linear combination of elements of S that equals zero; that is, for some nonempty subset $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset S$ we have*

$$a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_m \mathbf{x}_m = \mathbf{0},$$

where the elements $\mathbf{x}_i$ are distinct, and not all of the coefficients a_i are zero. If no such linear combination exists, then the set S is linearly independent.

Remark 1.2.9. The empty set is vacuously linearly independent. Because there are no vectors in $\emptyset$, there is no nontrivial linear combination of vectors.

Example 1.2.10. Let $V = \mathbb{F}^2$ and $S = \{\mathbf{x}, \mathbf{y}, \mathbf{z}\} \subset V$, where $\mathbf{x} = (1, 1)$, $\mathbf{y} = (-1, 1)$, and $\mathbf{z} = (0, 2)$. The set S is linearly dependent, since $\mathbf{x} + \mathbf{y} - \mathbf{z} = \mathbf{0}$.

Proposition 1.2.11. *If S is linearly independent, then any vector $\mathbf{v} \in \text{span}(S)$ can be written uniquely as a (finite) linear combination of elements of S . More precisely, for any distinct elements $\mathbf{x}_1, \dots, \mathbf{x}_m$ of S , if $\mathbf{v} = \sum_{i=1}^m a_i \mathbf{x}_i$ and $\mathbf{v} = \sum_{i=1}^m b_i \mathbf{x}_i$, then $a_i = b_i$ for every $i = 1, \dots, m$.*

Proof. Suppose that $\mathbf{v} = \sum_{i=1}^m a_i \mathbf{x}_i$ and $\mathbf{v} = \sum_{i=1}^m b_i \mathbf{x}_i$. Subtracting gives $\mathbf{0} = \sum_{i=1}^m (a_i - b_i) \mathbf{x}_i$. Since S is linearly independent, each term is equal to zero, which implies $a_i = b_i$ for each $i = 1, \dots, m$. $\square$

The next lemma is an important tool for two theorems in a later section (the replacement theorem (Theorem 1.4.1) and the extension theorem (Corollary 1.4.5)).

It states that linear independence is inherited by subsets. This is sort of a dual result to Proposition 1.2.5, which states that supersets of sets that span also span.

Lemma 1.2.12. *If S is linearly independent and $S' \subset S$, then S' is also linearly independent.*

Proof. See Exercise 1.13. $\square$

Lemma 1.2.12 and Proposition 1.2.5 suggest that linear independence and the span are complementary in some sense. There are some special sets that have both properties. These important sets are called *bases* of the vector space, and they act as a coordinate system for the vector space.

Definition 1.2.13. *If S is linearly independent and spans V , then it is called a basis of V .*

Example 1.2.14.

- (i) The vectors $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, and $\mathbf{e}_3 = (0, 0, 1)$ in $\mathbb{F}^3$ form a basis for $\mathbb{F}^3$.
- (ii) More generally the vectors $\mathbf{e}_1 = (1, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)$, $\dots$, $\mathbf{e}_n = (0, 0, \dots, 0, 1) \in \mathbb{F}^n$ form a basis of $\mathbb{F}^n$ called the *standard basis*.
- (iii) The vectors $(1, 1, 1)$, $(2, 1, 1)$, and $(1, 0, 1)$ also form a basis for $\mathbb{F}^3$. To show that the vectors are linearly independent, set

$$a(1, 1, 1) + b(2, 1, 1) + c(1, 0, 1) = (0, 0, 0).$$

Solving gives $a = b = c = 0$. To show that the vectors span $\mathbb{F}^3$, set

$$a(1, 1, 1) + b(2, 1, 1) + c(1, 0, 1) = (x, y, z).$$

Solving gives

$$a = y - x + z,$$

$$b = x - z,$$

$$c = z - y.$$

- (iv) The monomials $\{1, x, x^2, x^3, \dots\} \subset \mathbb{F}[x]$ form a basis of $\mathbb{F}[x]$ (the reader should prove this). But there are infinitely many other bases as well, for example, $\{1, x + 1, x^2 + x + 1, x^3 + x^2 + x + 1, \dots\}$.

Corollary 1.2.15. *If S is a basis for V , then any nonzero vector $\mathbf{x} \in V$ can be uniquely expressed as a linear combination of elements of S .*

Proof. Since S spans V , each $\mathbf{x} \in V$ can be expressed as a linear combination of elements of S . Uniqueness follows from Proposition 1.2.11. $\square$

This corollary means that if we have a basis $B = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of V , we can write any element $\mathbf{v} \in V$ uniquely as a linear combination $\mathbf{v} = \sum_{i=1}^n a_i \mathbf{x}_i$, and thus $\mathbf{v}$ can be identified with the n -tuple $(a_1, \dots, a_n) \in \mathbb{F}^n$. That is to say, the basis B has given us a *coordinate system* for V . A different basis would, of course, give us a different coordinate system. In the next chapter, we show how to transform vectors from one coordinate system to another.

1.3 Products, Sums, and Complements

In this section we describe two ways of constructing new vector spaces and subspaces from existing ones: the *Cartesian product* of vector spaces, and the *sum of subspaces* of a given vector space.

The Cartesian product of a finite collection of vector spaces (over the same field) is a vector space, where vector addition and scalar multiplication are defined componentwise. This allows us to combine vector spaces into a single larger vector space. For example, suppose we have two three-dimensional vector spaces, the first describing the position of a given particle, and the second describing the velocity of the particle. The Cartesian product of those two vector spaces forms a new, six-dimensional vector space describing both the position and velocity of the particle.

The sum of nontrivial subspaces of a given vector space is also a subspace. Moreover, if a pair of subspaces intersects only at zero, then their sum has no redundancy and is called a *direct sum*. A direct sum decomposes a subspace into the sum of other smaller subspaces. If a vector space is the direct sum of two subspaces, then we call these subspaces *complementary*. Decomposing a space into the “right” collection of complementary subspaces often simplifies a problem substantially.

1.3.1 Products

Proposition 1.3.1. *Let $V_1, V_2, \dots, V_n$ be a collection of vector spaces over the field $\mathbb{F}$. The Cartesian product*

$$V = \prod_{i=1}^n V_i = V_1 \times V_2 \times \dots \times V_n = \{(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) \mid \mathbf{v}_i \in V_i\}$$

forms a vector space over $\mathbb{F}$ with additive identity $(\mathbf{0}, \mathbf{0}, \dots, \mathbf{0})$, and vector addition and scalar multiplication defined componentwise as

- (i) $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) + (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) = (\mathbf{x}_1 + \mathbf{y}_1, \mathbf{x}_2 + \mathbf{y}_2, \dots, \mathbf{x}_n + \mathbf{y}_n),$
- (ii) $a(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = (a\mathbf{x}_1, a\mathbf{x}_2, \dots, a\mathbf{x}_n)$

for all $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n), (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \in V$ and $a \in \mathbb{F}$.

Proof. See Exercise 1.17. $\square$

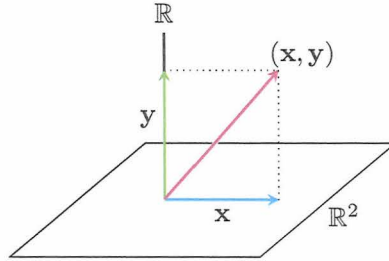


Figure 1.1. A depiction of $\mathbb{R}^2 \times \mathbb{R}$. Beware that $\mathbb{R}^2 \cup \mathbb{R}$ is not a vector space. The vector $(\mathbf{x}, \mathbf{y})$ in this diagram (red) lies in the product $\mathbb{R}^2 \times \mathbb{R}$, but it does not lie in $\mathbb{R}^2 \cup \mathbb{R}$.

Example 1.3.2. The product space $\prod_{i=1}^n \mathbb{F}$ is exactly the vector space $\mathbb{F}^n$.

Example 1.3.3. The space $\mathbb{R}^2 \times \mathbb{R}$ can be written as the set of pairs $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and $\mathbf{y} = y_1 \in \mathbb{R}$. The points of $\mathbb{R}^2 \times \mathbb{R}$ are in a natural bijective correspondence with the points of $\mathbb{R}^3$ by sending $((x_1, x_2), y_1)$ to (x_1, x_2, y_1) as in Figure 1.1.

Remark 1.3.4. It is straightforward to check that the previous proposition remains true for infinite sequences of vector spaces. For example, if $(V_i)_{i=1}^\infty$ is a sequence of vector spaces, the Cartesian product

$$\prod_{i=1}^{\infty} V_i = \{(\mathbf{x}_1, \mathbf{x}_2, \dots) \mid \mathbf{x}_i \in V_i \text{ for all } i = 1, 2, \dots\}$$

is a vector space with componentwise vector addition and scalar multiplication defined as follows:

- (i) $(\mathbf{x}_1, \mathbf{x}_2, \dots) + (\mathbf{y}_1, \mathbf{y}_2, \dots) = (\mathbf{x}_1 + \mathbf{y}_1, \mathbf{x}_2 + \mathbf{y}_2, \dots)$,
- (ii) $a(\mathbf{x}_1, \mathbf{x}_2, \dots) = (a\mathbf{x}_1, a\mathbf{x}_2, \dots)$.

1.3.2 Sums and Complements

Proposition 1.3.5. Let $W_1, W_2, \dots, W_n$ be a collection of subspaces of the vector space V . The sum $\sum_{i=1}^n W_i = \{\mathbf{w}_1 + \mathbf{w}_2 + \dots + \mathbf{w}_n \mid \mathbf{w}_i \in W_i\}$ is a subspace of V .

Proof. By Corollary 1.1.13 it suffices to show that $W = \sum_{i=1}^n W_i$ is nonempty and closed under vector addition and scalar multiplication. Since $\mathbf{0} \in W_i$ for every i , we have $\mathbf{0} \in \sum_{i=1}^n W_i$, so W is nonempty. If $\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n$ and $\mathbf{y}_1 + \mathbf{y}_2 + \dots + \mathbf{y}_n$ are in W , then so is their sum, since $(\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n) + (\mathbf{y}_1 + \mathbf{y}_2 + \dots + \mathbf{y}_n) = (\mathbf{x}_1 + \mathbf{y}_1) + (\mathbf{x}_2 + \mathbf{y}_2) + \dots + (\mathbf{x}_n + \mathbf{y}_n) \in W$. Moreover, we have that $a(\mathbf{w}_1 + \mathbf{w}_2 + \dots + \mathbf{w}_n) = a\mathbf{w}_1 + a\mathbf{w}_2 + \dots + a\mathbf{w}_n \in W$. $\square$

In the special case that subspaces W_1, W_2 satisfy $W_1 \cap W_2 = \{\mathbf{0}\}$, their sum behaves like a Cartesian product, as described in the following definition and theorem.

Definition 1.3.6. Let W_1, W_2 be subspaces of the vector space V . If $W_1 \cap W_2 = \{\mathbf{0}\}$, then the sum $W_1 + W_2$ is called a direct sum and is denoted $W_1 \oplus W_2$. More generally, if $(W_i)_{i=1}^n$ is a collection of subspaces of V , then the sum $W = \sum_{i=1}^n W_i$ is a direct sum if

$$W_i \cap \left(\sum_{j \neq i} W_j \right) = \{\mathbf{0}\} \quad \text{for all } i = 1, 2, \dots, n. \quad (1.2)$$

In this case we write $W = W_1 \oplus W_2 \oplus \dots \oplus W_n$ or $W = \bigoplus_{i=1}^n W_i$.

Theorem 1.3.7. Let $\{W_i\}_{i=1}^n$ be a collection of subspaces of a vector space V with bases $\{S_i\}_{i=1}^n$, respectively, and let $W = \sum_{i=1}^n W_i$. The following are equivalent:

- (i) $W = \bigoplus_{i=1}^n W_i$.
- (ii) For each $\mathbf{w} \in W$, there exists a unique n -tuple $(\mathbf{w}_1, \dots, \mathbf{w}_n)$, with each $\mathbf{w}_i \in W_i$, such that $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2 + \dots + \mathbf{w}_n$.
- (iii) $S = \bigcup_{i=1}^n S_i$ is a basis for W , and $S_i \cap S_j = \emptyset$ for every pair $i \neq j$.

Proof.

(i) $\Rightarrow$ (ii): If $\mathbf{w} = \mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n$ and $\mathbf{w} = \mathbf{y}_1 + \mathbf{y}_2 + \dots + \mathbf{y}_n$, each $\mathbf{x}_i, \mathbf{y}_i \in W_i$, then for each i , we have $\mathbf{x}_i - \mathbf{y}_i = \sum_{j \neq i} (\mathbf{y}_j - \mathbf{x}_j) \in W_i \cap \sum_{j \neq i} W_j = \{\mathbf{0}\}$. Hence, uniqueness holds.

(ii) $\Rightarrow$ (iii): Because each S_i is a basis of W_i , it is linearly independent; hence, $\mathbf{0} \notin S_i$. Since this holds for every i , we also have $\mathbf{0} \notin S$. If S is linearly dependent, then there exists some nontrivial linear combination of elements of S that equals zero. This contradicts the uniqueness of the representation in (ii), since the unique representation of zero is $\mathbf{0} = \mathbf{0} + \mathbf{0} + \dots + \mathbf{0}$.

Moreover, the set S spans W , since every element of W can be expressed as a linear combination of elements from $\{W_i\}_{i=1}^n$, which in turn can be expressed as a linear combination of elements of S_i .

Finally, if $\mathbf{s} \in S_i \cap S_j$ for some $i \neq j$, then the uniqueness of the representation in (ii), applied to $\mathbf{s} = \mathbf{s} + \mathbf{0} = \mathbf{0} + \mathbf{s}$, implies that $\mathbf{s} = \mathbf{0}$. But we have already seen that $\mathbf{0} \notin S$, so $S_i \cap S_j = \emptyset$.

(iii) $\Rightarrow$ (i): Suppose that for some i we have a nonzero element $\mathbf{w} \in W_i \cap \sum_{j \neq i} W_j = \text{span}\{S_i\} \cap \text{span}\{\bigcup_{j \neq i} S_j\}$. Thus, $\mathbf{w}$ is a nontrivial linear combination of elements of $S_i \subset S$ and a nontrivial linear combination of elements of $\bigcup_{j \neq i} S_j \subset S$. Since $S_i \cap (\bigcup_{j \neq i} S_j) = \emptyset$, this contradicts the uniqueness of linear combinations in the basis S . Thus, $W_i \cap \sum_{j \neq i} W_j = \{\mathbf{0}\}$ for all i . $\square$

Definition 1.3.8. Two subspaces W_1 and W_2 of V are complementary if $V = W_1 \oplus W_2$.

Example 1.3.9. Consider the space $\mathbb{F}[x; 2n]$. If $W_1 = \text{span}\{1, x^2, x^4, \dots, x^{2n}\}$ and $W_2 = \text{span}\{x, x^3, x^5, \dots, x^{2n-1}\}$, then $\mathbb{F}[x; 2n] = W_1 \oplus W_2$. In other words, W_1 and W_2 are complementary.

1.4 Dimension, Replacement, and Extension

In this section, we prove the replacement theorem, which says that if V has a finite basis of m elements, then a linearly independent subset of V has at most m elements. An immediate corollary of this is that if V has two bases and one of them is finite, then they both have the same cardinality. In other words, the number of elements in a given basis, if finite, is a meaningful constant, which we call the *dimension* of the vector space.

Another corollary of the replacement theorem is the extension theorem, which for a finite-dimensional vector space V , allows us to create a basis out of any linearly independent subset by adding vectors to the set. This is a powerful tool because it enables us to construct a basis from a certain subset of vectors that are desirable for the problem at hand.

We conclude this section with the difficult but important theorem that every vector space has a basis. The proof of this theorem uses Zorn's lemma, which is equivalent to the axiom of choice (see Appendix A for more details).

1.4.1 The Replacement and Extension Theorems

Theorem 1.4.1 (Replacement Theorem). *If V is spanned by $S = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$ and $T = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ (where each $\mathbf{t}_i$ is distinct) is a linearly independent subset of V , then $n \leq m$, and there exists a subset $S' \subset S$ having $m - n$ elements such that $T \cup S'$ spans V .*

Proof. We prove this by induction on n . If $n = 0$, then the result follows trivially (set $S' = S$). Now assume that the result holds for some $n \in \mathbb{N}$. It suffices to show that the result holds for $n + 1$. If $T = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{n+1}\}$ is a linearly independent subset of V , then we know that $T' = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ is also linearly independent by Lemma 1.2.12. Thus, by the inductive hypothesis, we have $n \leq m$, and there is a subset of S with $m - n$ elements, say $S' = \{\mathbf{s}_1, \dots, \mathbf{s}_{m-n}\}$, such that $T' \cup S'$ spans V . Hence, we can write $\mathbf{t}_{n+1}$ as a linear combination:

$$\mathbf{t}_{n+1} = a_1 \mathbf{t}_1 + \dots + a_n \mathbf{t}_n + b_1 \mathbf{s}_1 + \dots + b_{m-n} \mathbf{s}_{m-n}. \quad (1.3)$$

If $n = m$, then $S' = \emptyset$ and $\mathbf{t}_{n+1} \in \text{span}(T')$, which contradicts the linear independence of T . It follows that $n < m$ (hence $n + 1 \leq m$), and there exists at least one nonzero b_i in (1.3). Without loss of generality, assume $b_1 \neq 0$. Thus,

$$\mathbf{s}_1 = -\frac{1}{b_1} (a_1 \mathbf{t}_1 + \dots + a_n \mathbf{t}_n - \mathbf{t}_{n+1} + b_2 \mathbf{s}_2 + \dots + b_{m-n} \mathbf{s}_{m-n}). \quad (1.4)$$

It follows that $\mathbf{s}_1 \in \text{span}(T \cup S'')$, where $S'' = \{\mathbf{s}_2, \dots, \mathbf{s}_{m-n}\}$. By Proposition 1.2.7, we have that $\text{span}(T \cup S'') = \text{span}(T \cup S'' \cup \{\mathbf{s}_1\}) = \text{span}(T \cup S')$. But since $T' \cup S' \subset T \cup S'$, then by Proposition 1.2.5 we have $\text{span}(T \cup S') = V$. $\square$

Corollary 1.4.2. *If V has a basis of n elements, then all other bases of V have n elements.*

Proof. Let S be a basis of V having n elements. If T is any finite basis of V , then since S and T are both linearly independent and span V , the replacement theorem guarantees that both $|S| \leq |T|$ and $|T| \leq |S|$.

Suppose now that T is an infinite basis of V . Choose a subset $T' \subset T$ such that $|T'| = n + 1$. By Lemma 1.2.12 every subset of a linearly independent set is linearly independent, so by the replacement theorem, we have $|T'| \leq |S| = n$. This is a contradiction. $\square$

Definition 1.4.3. *A vector space V with a finite basis of n elements is n -dimensional or, equivalently, has dimension n . We denote this by $\dim(V) = n$. If V does not have a finite basis, then it is infinite dimensional.*

Corollary 1.4.4. *Let V be an n -dimensional vector space and $S \subset V$.*

- (i) *If S spans V , then S has at least n elements.*
- (ii) *If S is linearly independent, then S has at most n elements.*

Proof. These statements follow immediately from the replacement theorem. See Exercise 1.20. $\square$

Corollary 1.4.5 (Extension Theorem). *Let W be a subspace of the vector space V . If $S = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$ with each $\mathbf{s}_i$ distinct and $T = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ with each $\mathbf{t}_i$ distinct are bases for V and W , respectively, then there exists $S' \subset S$ having $m - n$ elements such that $T \cup S'$ is a basis for V .*

Proof. By the replacement theorem, there exists $S' \subset S$ such that $T \cup S'$ spans V . It suffices to show that $T \cup S'$ is linearly independent. Assume without loss of generality that $S' = \{\mathbf{s}_1, \dots, \mathbf{s}_{m-n}\}$ and suppose, by way of contradiction, there exists a nontrivial linear combination of elements of $T \cup S'$ satisfying

$$a_1 \mathbf{t}_1 + \dots + a_n \mathbf{t}_n + b_1 \mathbf{s}_1 + \dots + b_{m-n} \mathbf{s}_{m-n} = \mathbf{0}.$$

Since T is linearly independent, this implies that at least one of the b_i is nonzero. We assume without loss of generality that $b_1 \neq 0$. Thus, we have

$$\mathbf{s}_1 = -\frac{1}{b_1} (a_1 \mathbf{t}_1 + \dots + a_n \mathbf{t}_n + b_2 \mathbf{s}_2 + \dots + b_{m-n} \mathbf{s}_{m-n}).$$

Hence, $T \cup S''$ spans V , where $S'' = \{\mathbf{s}_2, \dots, \mathbf{s}_{m-n}\}$. This is a contradiction since $T \cup S''$ has only $m - 1$ elements, and Corollary 1.4.4 requires any set that spans the m -dimensional space V to have at least m elements. Thus, $T \cup S$ is linearly independent. $\square$

Example 1.4.6. Consider the linearly independent polynomials

$$f_1 = x^2 - 1, \quad f_2 = x^3 - x, \quad \text{and} \quad f_3 = x^3 - 2x^2 - x + 1$$

in the space $\mathbb{F}[x; 3]$. The set of monomials $\{1, x, x^2, x^3\}$ forms a basis for $\mathbb{F}[x; 3]$. Let $T = \{f_1, f_2, f_3\}$ and $S = \{1, x, x^2, x^3\}$. By the extension theorem there exists a subset S' of S such that $S' \cup T$ forms a basis for $\mathbb{F}[x; 3]$. One option is $S' = \{x\}$. A straightforward calculation shows

$$1 = f_2 - 2f_1 - f_3, \quad x^2 = f_2 - f_1 - f_3, \quad \text{and} \quad x^3 = f_2 + x,$$

so $\{f_1, f_2, f_3, x\}$ forms a basis for $\mathbb{F}[x; 3]$.

Corollary 1.4.7. *Let V be a finite-dimensional vector space. If W is a subspace of V and $\dim W = \dim V$, then $W = V$.*

Proof. This follows trivially from the extension theorem (Corollary 1.4.5). It is also proved in Exercise 1.21. $\square$

Vista 1.4.8. Both the replacement and extension theorems require that the underlying vector space be finite dimensional. This raises the question: Are infinite-dimensional vector spaces important? Or can we get away with narrowing our focus to the finite-dimensional case? Although you may have come pretty far in life just using finite-dimensional vector spaces, many important areas of mathematics rely heavily on infinite-dimensional vector spaces, such as $C([a, b]; \mathbb{F})$ and $L^\infty([a, b]; \mathbb{F})$. One example where such a vector space occurs is the study of differential equations, which includes most, if not all, of the laws of physics. Infinite-dimensional vector spaces are also widely used in other fields such as finance, economics, geology, climate science, biology, and nearly every area of engineering.

We conclude this section with a simple but useful result on dimension of direct sums and Cartesian products.

Proposition 1.4.9. *Let $\{W_i\}_{i=1}^n$ be a collection of finite-dimensional subspaces of the vector space V such that the direct-sum condition (1.2) holds. If $\dim(\bigoplus_{i=1}^n W_i)$ exists, then $\dim(\bigoplus_{i=1}^n W_i) = \sum_{i=1}^n \dim(W_i)$.*

Similarly, if $V_1, V_2, \dots, V_n$ is a collection of finite-dimensional vector spaces over the field $\mathbb{F}$, the Cartesian product

$$V = \prod_{i=1}^n V_i = V_1 \times V_2 \times \cdots \times V_n = \{(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) \mid \mathbf{v}_i \in V_i\}$$

satisfies

$$\dim(V_1 \times V_2 \times \cdots \times V_n) = \sum_{i=1}^n \dim(V_i).$$

Proof. The proof is Exercise 1.25. $\square$

1.4.2 * Every Vector Space Has a Basis

We now prove that every vector space has a basis—this is a major theorem in linear algebra. As an immediate consequence, this theorem says that the dimension of a vector space is well defined. Beyond that, an ordered basis is useful because it allows a vector to be represented as an array where each entry corresponds to the coefficient leading that basis element, that is, when the vector is written as a linear combination in that basis. By only keeping track of the array, we can think of the basis as a coordinate system for a vector space. Thus, in addition to showing that a basis exists, this theorem also tells that we always have a coordinate system for a given vector space.

The proof follows from Zorn's lemma, which is equivalent to the axiom of choice.⁵ We take these as axioms of set theory; for more details, see the discussion in Appendix A.4.

Theorem 1.4.10 (Zorn's Lemma). *Let $(X, \leq)$ be a nonempty partially ordered set.⁶ If every chain in X has an upper bound in X , then X contains a maximal element.*

By *chain* we mean any subset $C \subset X$ such that C is totally ordered; that is, for every $\alpha, \beta \in C$ we have either $\alpha \leq \beta$ or $\beta \leq \alpha$. A chain C is said to have an *upper bound* in X if there is an element $\gamma \in X$ such that $\alpha \leq \gamma$ for every $\alpha \in C$.

Theorem 1.4.11. *Every vector space has a basis.*

Although this theorem is about bases of vector spaces, the idea of its proof is useful in many settings where one needs to prove that a maximal set having certain properties exists.

Proof. Let V be any vector space. If $V = \{0\}$, then the empty set is a basis. Hence, we may assume that V is nontrivial.

Let $\mathcal{S} = \{S_\alpha\}_{\alpha \in I}$ be the set of all linearly independent subsets S_α of V . Since V is nontrivial, $\mathcal{S}$ is not empty. The set $\mathcal{S}$ is partially ordered by set inclusion $\subset$. To use Zorn's lemma, we must show that every chain in $\mathcal{S}$ has an upper bound.

⁵Proofs that rely on the axiom of choice are usually only proofs of existence. Specifically, the theorem here about existence of a basis doesn't say anything about how to construct a basis.

⁶See Definition A.3.15.

A nontrivial chain in $\mathcal{S}$ is a nonempty subset $C = \{S_\alpha\}_{\alpha \in J} \subset \mathcal{S}$ that is totally ordered by set inclusion. We claim that the set $S' = \bigcup_{\alpha \in J} S_\alpha$ is an upper bound for C ; that is, $S' \in \mathcal{S}$ and $S_\alpha \subset S'$ for all $\alpha \in J$. The inclusion $S_\alpha \subset S'$ is immediate, so we need only show that S' is linearly independent.

If S' were not linearly independent, then there would exist a nontrivial linear combination $a_1 \mathbf{x}_1 + \cdots + a_m \mathbf{x}_m = \mathbf{0}$ where $\mathbf{x}_i \in S'$. By definition of S' , we have, for each $1 \leq i \leq m$, that there is some $\alpha_i \in J$ such that $\mathbf{x}_i \in S_{\alpha_i}$. Without loss of generality assume that $S_{\alpha_1} \subset S_{\alpha_2} \subset \cdots \subset S_{\alpha_m}$. Hence, $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset S_{\alpha_m}$, which is a linearly independent set. This is a contradiction. Therefore, S' must be linearly independent, and every chain in $\mathcal{S}$ has an upper bound.

By Zorn's lemma, there exists a maximal element $S \in \mathcal{S}$. We claim that S is a basis for V . It suffices to show that S spans V . If not, then there exists $\mathbf{v} \in V$ which is not in the span of S . Thus, $\{\mathbf{v}\} \cup S$ is linearly independent and properly contains S , which contradicts the maximality of S . Hence, S must span V , as desired, and is therefore a basis for V . $\square$

1.5 Quotient Spaces

This section describes yet another way to construct a vector space from a subspace. Specifically, we form a vector space by considering the set of all the translates of a given subspace. There is a natural way to define vector addition and scalar multiplication of these translates that makes the set of translates into a vector space called the *quotient space*.

As an analogy, it may be helpful to think of quotient spaces as a vector-space analogue of modular arithmetic. Recall that in modular arithmetic we start with a set S of all integers divisible by a certain number. So, for example, we could let $S = 5\mathbb{Z} = \{\dots, -15, -10, -5, 0, 5, 10, \dots\}$ be the set of numbers divisible by 5. The *translates* of S are the sets

$$\begin{aligned} 0 + S &= S = \{\dots, -15, -10, -5, 0, 5, 10, \dots\}, \\ 1 + S &= \{\dots, -14, -9, -4, 1, 6, 11, \dots\}, \\ 2 + S &= \{\dots, -13, -8, -3, 2, 7, 12, \dots\}, \\ 3 + S &= \{\dots, -12, -7, -2, 3, 8, 13, \dots\}, \\ 4 + S &= \{\dots, -11, -6, -1, 4, 9, 14, \dots\}. \end{aligned}$$

The set of these translates is usually written $\mathbb{Z}/5\mathbb{Z}$, and it has only 5 elements in it:

$$\mathbb{Z}/5\mathbb{Z} = \{0 + S, 1 + S, 2 + S, 3 + S, 4 + S\}.$$

There are only 5 translates of S because $0 + S = 5 + S = 10 + S = \dots$, and $2 + S = 12 + S = -8 + S = \dots$, and so forth. When we say *6 is equivalent to 1 modulo 5* or write $6 \equiv 1 \pmod{5}$, we mean that the number 6 lies in the translate $1 + S$, or equivalently $6 - 1 \in S$, or equivalently $1 + S = 6 + S$.

It is a very useful fact that addition, subtraction, and multiplication all make sense modulo S ; that is, we can add, subtract, or multiply the translates by adding, subtracting, or multiplying any of their corresponding elements. So when we write $3 + 4 \equiv 2 \pmod{5}$ we mean that if we take any element in $3 + S$ and add it to any element in $4 + S$, we get an element in $2 + S$.

The construction of a vector-space quotient is very similar. Instead of a set of multiples of a given number, we take a subspace W of a given vector space V . The translates of W are sets of the form $\mathbf{x} + W = \{\mathbf{x} + \mathbf{w} \mid \mathbf{w} \in W\}$, where $\mathbf{x} \in V$. The set of all translates is written V/W and is called the *quotient of V by W* . It follows that $\mathbf{x}_1 + W = \mathbf{x}_2 + W$ if and only if $\mathbf{x}_1 - \mathbf{x}_2 \in W$. Thus, we say that $\mathbf{x}_1$ is equivalent to $\mathbf{x}_2$ if $\mathbf{x}_1 + W = \mathbf{x}_2 + W$, or equivalently if $\mathbf{x}_1 - \mathbf{x}_2 \in W$.

Quotient spaces provide useful tools for better understanding several important concepts, including linear transformations, systems of linear equations, Lebesgue integration, and the Chinese remainder theorem. The key fact that makes these quotient spaces so useful is the fact that the set V/W of these translates is itself a vector space. That is, we can add, subtract, and scalar multiply the translates, in direct analogy to what we did in modular arithmetic.

In this section, we define quotient spaces and the corresponding operations of vector addition and scalar multiplication on quotient spaces. We also give several examples.

Throughout this section, we assume that W is a subspace of a given vector space V .

1.5.1 Cosets

Definition 1.5.1. We say that $\mathbf{x}$ is equivalent to $\mathbf{y}$ modulo W if $\mathbf{x} - \mathbf{y} \in W$; this is denoted $\mathbf{x} \sim_W \mathbf{y}$ or just by $\mathbf{x} \sim \mathbf{y}$.

Proposition 1.5.2. The relation $\sim$ is an equivalence relation.⁷

Proof. It suffices to show that the relation $\sim$ is (i) reflexive, (ii) symmetric, and (iii) transitive.

- (i) Note that $\mathbf{x} - \mathbf{x} = \mathbf{0} \in W$. Thus, $\mathbf{x} \sim \mathbf{x}$.
- (ii) If $\mathbf{x} - \mathbf{y} \in W$, then $\mathbf{y} - \mathbf{x} = (-1)(\mathbf{x} - \mathbf{y}) \in W$. Hence, $\mathbf{x} \sim \mathbf{y}$ implies $\mathbf{y} \sim \mathbf{x}$.
- (iii) If $\mathbf{x} - \mathbf{y} \in W$ and $\mathbf{y} - \mathbf{z} \in W$, then $\mathbf{x} - \mathbf{z} = (\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{z}) \in W$. Thus, $\mathbf{x} \sim \mathbf{y}$ and $\mathbf{y} \sim \mathbf{z}$ implies $\mathbf{x} \sim \mathbf{z}$. $\square$

Recall that an equivalence relation defines a *partition*. More precisely, for each $\mathbf{x} \in V$, we define the *equivalence class of $\mathbf{x}$* to be the set $[\mathbf{x}] = \{\mathbf{y} \mid \mathbf{y} \sim \mathbf{x}\}$. Thus, every element $\mathbf{x} \in V$ is in exactly one equivalence class.

The equivalence classes defined by $\sim$ have a particularly useful structure. Note that

$$\begin{aligned} \{\mathbf{y} \mid \mathbf{y} \sim \mathbf{x}\} &= \{\mathbf{y} \mid \mathbf{y} - \mathbf{x} \in W\} \\ &= \{\mathbf{y} \mid \mathbf{y} = \mathbf{x} + \mathbf{w} \text{ for some } \mathbf{w} \in W\} \\ &= \{\mathbf{x} + \mathbf{w} \mid \mathbf{w} \in W\}. \end{aligned}$$

⁷The definition of an equivalence relation is given in Appendix A.1.

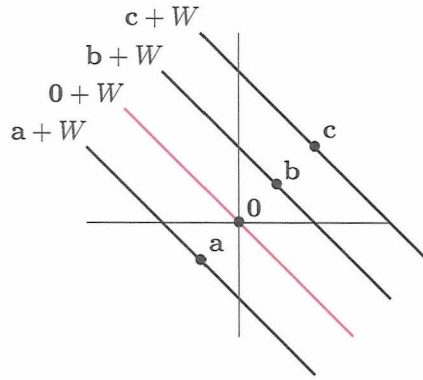


Figure 1.2. A representation of the cosets of a subspace W as translates of W . Here W (red) is the subspace $\text{span}(\{(1, -1)\}) = \{(x, y) \mid y = -x\}$, and each of the other lines represents a coset of W . If $\mathbf{b} = (b_1, b_2)$, then the coset $\mathbf{b} + W$ is the line $\{\mathbf{b} + (x, y) \mid y = -x\} = \{(x + b_1, y + b_2) \mid y = -x\}$.

Hence, the equivalence classes of $\sim$ are just translates of the subspace W , and we often denote them by $\mathbf{x} + W = \{\mathbf{x} + \mathbf{w} \mid \mathbf{w} \in W\} = [\mathbf{x}]$; see Figure 1.2. The equivalence classes of W are called the *cosets* of W . Cosets are either identical or disjoint, and we have $\mathbf{x} + W = \mathbf{x}' + W$ if and only if $\mathbf{x} - \mathbf{x}' \in W$. As we show in Theorem 1.5.7 below, the set of all cosets of W is a vector space.

1.5.2 Quotient Vector Spaces

Definition 1.5.3. The set $\{\mathbf{x} + W \mid \mathbf{x} \in V\}$ (or equivalently $\{[\mathbf{x}] \mid \mathbf{x} \in V\}$) of all cosets of W in V is denoted V/W and is called the *quotient of V modulo W* .

Example 1.5.4.

- (i) Let $V = \mathbb{R}^3$ and let $W = \text{span}((0, 1, 0))$ be the y -axis. We show that there is a natural bijective correspondence between the elements (cosets) of the quotient V/W and the elements of $\mathbb{R}^2$.

Note that any $(a, b, c) \in V$ can be written as $(a, b, c) = (a, 0, c) + (0, b, 0)$, and $(0, b, 0) \in W$. Therefore, the coset $(a, b, c) + W$ in the quotient V/W is equal to the coset $(a, 0, c) + W$, and we have a surjection $\varphi : \mathbb{R}^2 \rightarrow V/W$, defined by sending $(a, c) \in \mathbb{R}^2$ to $(a, 0, c) + W \in V/W$.

If $\varphi(a, c) \sim \varphi(a', c')$, then $(a, 0, c) \sim (a', 0, c')$, implying that $(a, 0, c) - (a', 0, c') = (a - a', 0, c - c') \in W = \{(0, b, 0) \mid b \in \mathbb{R}\}$. It follows that $a - a' = 0 = c - c'$, and so $a = a'$ and $c = c'$. Thus, the map φ is injective. This gives a bijection from $\mathbb{R}^2$ to V/W . Below we show that V/W has a natural vector-space structure, and the bijection φ preserves all the properties of a vector space.

- (ii) Let $V = C([0, 1]; \mathbb{R})$ be the vector space of real-valued functions defined and continuous on the interval $[0, 1]$, and let $W = \{f \in V \mid f(0) = 0\}$ be the subspace of all functions that vanish at 0. We show that there is a natural bijective correspondence between the elements (cosets) of V/W and the real numbers.

Given any function $f \in V$, let $\bar{f}(x) = f(x) - f(0) \in W$, and let f_0 be the constant function $f_0(x) = f(0)$. We can check that $\bar{f} \in W$ and that $f_0 \in V$. Thus, $f = f_0 + \bar{f}$. This shows that $f(x) \sim f_0(x)$ modulo W , and the coset $f + W$ in V/W can be written as $f_0 + W$.

Now we can proceed as in the previous example. Given any $a \in \mathbb{R}$ there is a corresponding constant function in V , which we also denote by $f_a(x) = a$. Let $\psi : \mathbb{R} \rightarrow V/W$ be defined as $\psi(a) = f_a + W$. This map is surjective, since any $f + W \in V/W$ can be written as $f_0 + W = \psi(f(0))$. Also, given any $a, a' \in \mathbb{R}$, if $\psi(a) \sim \psi(a')$, then $f_a - f_{a'} \in W$. But $f_a - f_{a'}$ is constant, and since it vanishes at 0, it must vanish everywhere; that is, $a = a'$. It follows that ψ is injective and thus also a bijection.

- (iii) Let $V = \mathbb{F}[x]$ and let $W = \text{span}(\{x^3, x^4, \dots\})$ be the subspace of $\mathbb{F}[x]$ consisting of all polynomials with no nonzero terms of degree less than 3. Any $f = a_0 + a_1x + \dots + a_nx^n \in \mathbb{F}[x]$ is equivalent mod W to $a_0 + a_1x + a_2x^2$, and so the coset $f + W$ can be written as $a_0 + a_1x + a_2x^2 + W$. An argument similar to the previous example shows that the quotient V/W is in bijective correspondence with the set $\{(a_0, a_1, a_2) \mid a_i \in \mathbb{F}\} = \mathbb{F}^3$ via the map $(a_0, a_1, a_2) \mapsto a_0 + a_1x + a_2x^2 + W$.

Definition 1.5.5. *The operations of vector addition $\boxplus : V/W \times V/W \rightarrow V/W$ and scalar multiplication $\boxtimes : \mathbb{F} \times V/W \rightarrow V/W$ on the quotient space V/W are defined to be*

- (i) $(\mathbf{x} + W) \boxplus (\mathbf{y} + W) = (\mathbf{x} + \mathbf{y}) + W$, and
(ii) $a \boxtimes (\mathbf{x} + W) = (a\mathbf{x}) + W$.

Lemma 1.5.6. *The operations $\boxplus : V/W \times V/W \rightarrow V/W$ and $\boxtimes : \mathbb{F} \times V/W \rightarrow V/W$ are well defined for all $\mathbf{x}, \mathbf{y} \in V$ and $a \in \mathbb{F}$.*

Figure 1.3 shows the additive part of this lemma in $\mathbb{R}^2$: different representatives of each coset can sum to different vectors, but those different sums still lie in the same coset.

Proof.

- (i) We must show that the definition of $\boxplus$ does not depend on the choice of representative of the coset; that is, if $\mathbf{x} + W = \mathbf{x}' + W$ and $\mathbf{y} + W = \mathbf{y}' + W$, then we must show $(\mathbf{x} + W) \boxplus (\mathbf{y} + W) = (\mathbf{x}' + W) \boxplus (\mathbf{y}' + W)$.

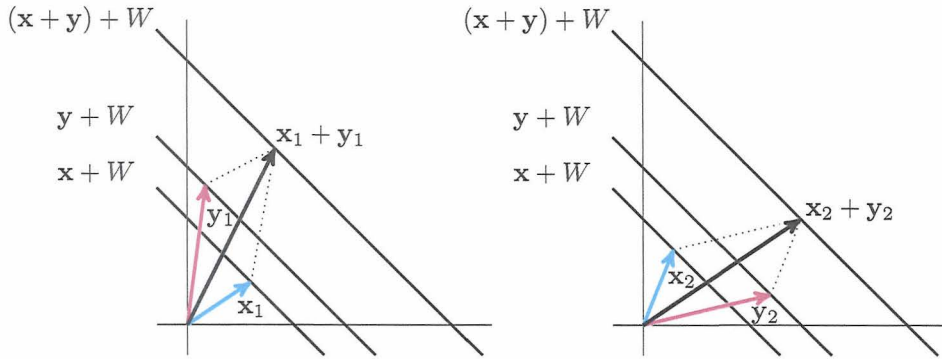


Figure 1.3. For any two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in the coset $\mathbf{x} + W$ and any other two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ in the coset $\mathbf{y} + W$, the sums $\mathbf{x}_1 + \mathbf{y}_1$ and $\mathbf{x}_2 + \mathbf{y}_2$ both lie in the coset $(\mathbf{x} + \mathbf{y}) + W$.

Because $\mathbf{x} + W = \mathbf{x}' + W$, we have that $\mathbf{x} - \mathbf{x}' \in W$. And similarly, because $\mathbf{y} + W = \mathbf{y}' + W$, we have $\mathbf{y} - \mathbf{y}' \in W$. Adding yields $(\mathbf{x} - \mathbf{x}') + (\mathbf{y} - \mathbf{y}') \in W$, which can be rewritten as $(\mathbf{x} + \mathbf{y}) - (\mathbf{x}' + \mathbf{y}') \in W$. Thus, we have that $(\mathbf{x} + \mathbf{y}) + W = (\mathbf{x}' + \mathbf{y}') + W$, and so $(\mathbf{x} + W) \boxplus (\mathbf{y} + W) = (\mathbf{x}' + W) \boxplus (\mathbf{y}' + W)$.

- (ii) Similarly, we must show that the operation $\boxtimes$ does not depend on the choice of representative of the coset; that is, if $\mathbf{x} + W = \mathbf{x}' + W$ and $a \in \mathbb{F}$, we show that $a \boxtimes (\mathbf{x} + W) = a \boxtimes (\mathbf{x}' + W)$. Since $\mathbf{x} + W = \mathbf{x}' + W$, we have that $\mathbf{x} - \mathbf{x}' \in W$, which implies that $a(\mathbf{x} - \mathbf{x}') \in W$, or equivalently $a\mathbf{x} - a\mathbf{x}' \in W$. Thus, we see $a\mathbf{x} + W = a\mathbf{x}' + W$. $\square$

Theorem 1.5.7. The quotient space V/W is a vector space when endowed with the operations $\boxplus$ and $\boxtimes$ of vector addition and scalar multiplication, respectively.

Proof. The operation $\boxplus$ is commutative because $+$ is commutative. That is, for any cosets $(\mathbf{x} + W) \in V/W$ and $(\mathbf{y} + W) \in V/W$, we have $(\mathbf{x} + W) \boxplus (\mathbf{y} + W) = (\mathbf{x} + \mathbf{y}) + W = (\mathbf{y} + \mathbf{x}) + W = (\mathbf{y} + W) \boxplus (\mathbf{x} + W)$. The proof of associativity is similar.

We note that the additive identity in V/W is $\mathbf{0} + W$, since for any $\mathbf{x} + W \in V/W$ we have $(\mathbf{0} + W) \boxplus (\mathbf{x} + W) = (\mathbf{0} + \mathbf{x}) + W = \mathbf{x} + W$. Similarly, the additive inverse of the coset $\mathbf{x} + W$ is $(-\mathbf{x}) + W$, since $(\mathbf{x} + W) \boxplus ((-\mathbf{x}) + W) = (\mathbf{x} + (-\mathbf{x})) + W = \mathbf{0} + W$.

The remaining details are left for the reader; see Exercise 1.26. $\square$

Example 1.5.8.

- (i) Consider again the example of $V = \mathbb{R}^3$ and $W = \text{span}((0, 1, 0))$, as in Example 1.5.4(i). We saw already that the map φ is a bijection from $\mathbb{R}^2$ to V/W , but now V/W also has a vector-space structure. We now show that φ has the special property that it preserves both vector addition and scalar multiplication. (Note: Not every bijection from $\mathbb{R}^2$ to V/W has this property.)

The operation $\boxplus$ is given by $((a, b, c) + W) \boxplus ((a', b', c') + W) = (a + a', b + b', c + c') + W$. The map φ satisfies $\varphi(a, c) \boxplus \varphi(a', c') = ((a, 0, c) + W) \boxplus ((a', 0, c') + W) = (a + a', 0, c + c') + W = \varphi((a, c) + (a', c'))$.

Similarly, for any scalar $d \in \mathbb{R}$ we have $d \boxtimes \varphi(a, c) = d \boxtimes (a, 0, c) + W = (da, 0, dc) + W = \varphi(d(a, c))$.

- (ii) In the case of $V = C([0, 1]; \mathbb{R})$ and $W = \{f \in V \mid f(0) = 0\}$, we have already seen in Example 1.5.4(ii) that any coset can be uniquely written as $a + W$, with $a \in \mathbb{R}$, giving the bijection $\psi : \mathbb{R} \rightarrow V/W$. Again the map ψ is especially nice because for any $a, b \in \mathbb{R}$ we have $\psi((a+W) \boxplus (b+W)) = \psi((a+b)+W) = a+b = \psi(a+W) + \psi(b+W)$, and for any $d \in \mathbb{R}$ we have $\psi(d \boxtimes (a+W)) = \psi(da+W) = da = d\psi(a+W)$.
- (iii) In the case of $V = \mathbb{F}[x]$ and $W = \text{span}(\{x^3, x^4, \dots\})$, any two cosets can be written in the form $(a_0 + a_1x + a_2x^2) + W$ and $(b_0 + b_1x + b_2x^2) + W$, and the operation $\boxplus$ is given by $((a_0 + a_1x + a_2x^2) + W) \boxplus ((b_0 + b_1x + b_2x^2) + W) = ((a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2) + W$. Similarly, the operation of $\boxtimes$ is given by $d \boxtimes ((a_0 + a_1x + a_2x^2) + W) = (da_0 + da_1x + da_2x^2) + W$.

Example 1.5.9.

- (i) In the case of $V = \mathbb{R}^3$ and $W = \text{span}((0, 1, 0))$, we saw in the previous example that the operations of $\boxplus$ and $\boxtimes$ on the space $\mathbb{R}^3/W$ are identical (via φ) to the usual operations of vector addition $+$ and scalar multiplication $\cdot$ on the vector space $\mathbb{R}^2$. So the vector space V/W and the vector space $\mathbb{R}^2$ are equivalent or “the same” in a sense that is defined more precisely in the next chapter.
- (ii) Similarly, in the case of $V = C([0, 1]; \mathbb{R})$ and $W = \{f \in V \mid f(0) = 0\}$ the map ψ preserves vector addition and scalar multiplication, so the space $C([0, 1]; \mathbb{R})/W$ is “the same as” the vector space $\mathbb{R}$.
- (iii) Finally, in the case that $V = \mathbb{F}[x]$ and $W = \text{span}(\{x^3, x^4, \dots\})$, the space $\mathbb{F}[x]/W$ is “the same as” the vector space $\mathbb{F}^3$. Of course, these are different sets, and they have other structure beyond the vector addition and scalar multiplication, but *as vector spaces* they are the same.

Vista 1.5.10 (Quotient Spaces and Integration). Quotient spaces play an important role in the theory of integration. If two functions are the same on all but a very small set, they have the same integral. For example, the function

$$h(x) = \begin{cases} 1, & x \neq 1/2, \\ 9, & x = 1/2, \end{cases}$$

has integral 1 on the interval $[0, 1]$. Similarly the function

$$f(x) = 1$$

also has integral 1 on the interval $[0, 1]$. When studying integration, it is helpful to think of these two functions as being the same.

A set Z has *measure zero* if it is so small that any two functions that differ only on Z must have the same integral. The set of all integrable functions that are supported only on a set of measure zero is a subspace of the space of all integrable functions.

The precise way to make sense of the idea that functions like f and h should be “the same” is to work with the quotient of the space of all integrable functions by the subspace of functions supported on a set of measure zero. We treat these ideas much more carefully in Chapters 8 and 9.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 1.1. Show that the set $V = (0, \infty)$ is a vector space over $\mathbb{R}$ with vector addition ($\oplus$) and scalar multiplication ($\odot$) defined as

$$x \oplus y = xy$$

$$a \odot x = x^a,$$

where $a \in \mathbb{R}$ and $x, y \in V$.

- 1.2. Finish the proof of Theorem 1.1.12 by proving properties (iii) and (iv). Also explain why the proof of the multiplicative identity property (vii) follows immediately from the fact that V is a vector space, but the additive identity property (iii) does not.

1.3. Do the following sets form subspaces of $\mathbb{F}^3$? Prove or disprove.

- (i) $\{(x_1, x_2, x_3) \mid 3x_1 + 4x_3 = 1\}$.
- (ii) $\{(x_1, x_2, x_3) \mid x_1 = 2x_2 = x_3\}$.
- (iii) $\{(x_1, x_2, x_3) \mid x_3 = x_1 + 4x_2\}$.
- (iv) $\{(x_1, x_2, x_3) \mid x_3 = 2x_1 \text{ or } x_3 = 3x_2\}$.

1.4. Do the following sets form subspaces of $\mathbb{F}[x; 4]$? (See Example 1.1.15(iv).) Prove or disprove.

- (i) The set of polynomials in $\mathbb{F}[x; 4]$ of even degree.
- (ii) The set of all polynomials of degree 3.
- (iii) The set of all polynomials $p(x)$ in $\mathbb{F}[x; 4]$ such that $p(0) = 0$.
- (iv) The set of all polynomials $p(x)$ in $\mathbb{F}[x; 4]$ such that $p'(0) = 0$.
- (v) The set of all polynomials in $\mathbb{F}[x; 4]$ having at least one real root.

1.5. Assume that $1 \leq p < q < \infty$. Is it true that ℓ^p is a subspace of ℓ^q ? (See Example 1.1.6(vi).) Justify your answer.

1.6. Let $\mathbf{x}_1 = (-1, 2, 3)$ and $\mathbf{x}_2 = (3, 4, 2)$.

- (i) Is $\mathbf{x} = (2, 6, 6)$ in $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2\})$? Justify your answer.
- (ii) Is $\mathbf{y} = (-9, -2, 5)$ in $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2\})$? Justify your answer.

1.7. For each set below, which of the following span $\mathbb{F}[x; 2]$? Justify your answer.

- (i) $\{1, x^2, x^2 - 3\}$.
- (ii) $\{1, x - 1, x^2 - 1\}$.
- (iii) $\{x + 2, x - 2, x^2 - 2\}$.
- (iv) $\{x + 4, x^2 - 4\}$.
- (v) $\{x - 1, x + 1, 2x + 3, 5x - 7\}$.

1.8. Which of the sets in Exercise 1.7 are linearly independent? Justify your answer.

1.9. Assume that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ spans V , and let $\mathbf{v}$ be any other vector in V . Show that the set $\{\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is linearly dependent.

1.10. Prove Proposition 1.2.5.

1.11. Prove that the monomials $\{1, x, x^2, x^3, \dots\} \subset \mathbb{F}[x]$ form a basis of $\mathbb{F}[x]$.

1.12. Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a linearly independent set of distinct vectors in V . Show that $\{\mathbf{v}_2, \dots, \mathbf{v}_n\}$ does not span V .

1.13. Prove Lemma 1.2.12.

1.14. Prove that the converse of Corollary 1.2.15 is also true; that is, if any vector in V can be uniquely expressed as a linear combination of elements of a set S , then S is a basis.

-
- 1.15. Prove that there is a basis for $\mathbb{F}[x; n]$ consisting only of polynomials of degree $n \in \mathbb{N}$.
- 1.16. For every positive integer n , write $\mathbb{F}[x]$ as the direct sum of n subspaces.
- 1.17. Prove Proposition 1.3.1.
- 1.18. Let

$$\begin{aligned}\text{Sym}_n(\mathbb{F}) &= \{A \in M_n(\mathbb{F}) \mid A^\top = A\}, \\ \text{Skew}_n(\mathbb{F}) &= \{A \in M_n(\mathbb{F}) \mid A^\top = -A\}.\end{aligned}$$

Show that

- (i) $\text{Sym}_n(\mathbb{F})$ is a subspace of $M_n(\mathbb{F})$,
 - (ii) $\text{Skew}_n(\mathbb{F})$ is a subspace of $M_n(\mathbb{F})$, and
 - (iii) $M_n(\mathbb{F}) = \text{Sym}_n(\mathbb{F}) \oplus \text{Skew}_n(\mathbb{F})$.
- 1.19. Show that any function $f \in C([-1, 1]; \mathbb{R})$ can be uniquely expressed as an even continuous function on $[-1, 1]$ plus an odd continuous function on $[-1, 1]$. Then show that the spaces of even and odd continuous functions of $C([-1, 1]; \mathbb{R})$, respectively, form complementary subspaces of $C([-1, 1]; \mathbb{R})$.
-
- 1.20. Prove Corollary 1.4.4.
- 1.21. Let W be a subspace of the finite-dimensional vector space V . Prove, not using any results after Section 1.4, that if $\dim(W) = \dim(V)$, then $W = V$. Hint: Prove the contrapositive.
- 1.22. Let W be a subspace of the n -dimensional vector space V . Prove that $\dim(W) \leq n$.
- 1.23. Prove: Let V be an n -dimensional vector space. If $S \subset V$ is a linearly independent subset of V with n elements, then S is a basis.
- 1.24. Let W be a subspace of the finite-dimensional vector space V . Use the extension theorem (Corollary 1.4.5) to show that there is a subspace X such that $V = W \oplus X$.
- 1.25. Prove Proposition 1.4.9. Hint: Consider using Theorem 1.3.7.

-
- 1.26. Using the operations defined in Definition 1.5.5, prove the remaining details in Theorem 1.5.7.
- 1.27. Prove that the quotient V/W satisfies

$$(a \sqcap (\mathbf{x} + W)) \boxplus (b \sqcap (\mathbf{y} + W)) = (a\mathbf{x} + b\mathbf{y}) + W.$$

- 1.28. Show that the quotient V/V consists of a single element.
- 1.29. Show that the quotient $V/\{\mathbf{0}\}$ has an obvious bijection $\phi: V/\{\mathbf{0}\} \rightarrow V$ which satisfies $\phi(\mathbf{x} + \{\mathbf{0}\}) \boxplus \phi(\mathbf{y} + \{\mathbf{0}\}) = \phi(\mathbf{x} + \{\mathbf{0}\}) + \phi(\mathbf{y} + \{\mathbf{0}\})$ and $\phi(c \sqcap (\mathbf{x} + \{\mathbf{0}\})) = c\phi(\mathbf{x} + \{\mathbf{0}\})$ for all $\mathbf{x}, \mathbf{y} \in V$ and $c \in \mathbb{F}$.

- 1.30. Let $V = \mathbb{F}[x]$ and $W = \text{span}(\{x, x^3, x^5, \dots\})$ be the span of the set of all odd-degree monomials. Prove that there is a bijective map $\psi : V/W \rightarrow \mathbb{F}[y]$ satisfying $\psi(p + W \oplus q + W) = \psi(p + W) + \psi(q + W)$ and $\psi(c \cdot (p + W)) = c\psi(p + W)$.

Notes

For a friendly description of many modern applications of linear algebra, we recommend Tim Chartier's little book [Cha15].

The reader who wants to review elementary linear algebra may find some of the following books useful [Lay02, Leo80, OS06, Str80]. G. Strang also has some very clear video explanations of many important linear algebra concepts available through the MIT Open Courseware page [Str10].

2

Linear Transformations and Matrices

The Matrix is everywhere. It is all around us. Even now, in this very room. You can see it when you look out your window, or when you turn on your television. You can feel it when you go to work, when you go to church, when you pay your taxes. It is the world that has been pulled over your eyes to blind you from the truth.

—Morpheus

A *linear transformation* (or linear map) is a function between two vector spaces that preserves all the linear structures, that is, lines map into lines, the origin maps to the origin, and subspaces map into subspaces.

The study of linear transformations of vector spaces can be broken down into roughly three areas, namely, algebraic properties, geometric properties, and operator-theoretic properties. In this chapter, we explore the algebraic properties of linear transformations. We discuss the geometric properties in Chapter 3 and the operator-theoretic properties in Chapter 4.

We begin by defining linear transformations and describing their attributes. For example, when a linear transformation has an inverse, we say that it is an *isomorphism*, and that the domain and codomain are *isomorphic*. This is a mathematical way of saying that, as vector spaces, the domain and codomain are the same. One of the big results in this chapter is that all n -dimensional vector spaces over the field $\mathbb{F}$ are isomorphic to $\mathbb{F}^n$.

We also consider basic features of linear transformations such as the kernel and range. Of particular interest is the first isomorphism theorem, which states that the quotient space of the domain of a linear transformation, modulo its kernel, is isomorphic to its range. This is used to prove several results about the dimensions of various spaces. In particular, an important consequence of the first isomorphism theorem is the rank-nullity theorem, which tells us that for a given linear map, the dimension of the domain is equal to the dimension of its range (called rank) plus the dimension of the kernel (called nullity). The rank-nullity theorem is a major result in linear algebra and is frequently used in applications.

Another useful consequence of the first isomorphism theorem is the second isomorphism theorem, which provides an identity involving sums and intersections

of subspaces. Although it may appear to be of little importance initially, the second isomorphism theorem is used to prove the *dimension formula* for subspaces, which says that the dimension of the sum of two subspaces is equal to the sum of the dimensions of each subspace minus the dimension of their intersection. This is an intuitive counting formula that we use later on.

An important branch of linear algebra is matrix theory. Matrices appear as representations of linear transformations between two finite-dimensional spaces, where matrix-vector multiplication represents the mapping of a vector by the linear transformation. This matrix representation is unique once a basis for the domain and a basis for the codomain are chosen. This raises the question of how the matrix representation of a linear transformation changes when the underlying bases change. For each linear transformation, there are certain canonical bases that are natural to use.

In an elementary linear algebra class, a great deal of attention is given to solving linear systems; that is, given an $m \times n$ matrix A and the vector $\mathbf{b} \in \mathbb{F}^m$, find the vector $\mathbf{x} \in \mathbb{F}^n$ satisfying $A\mathbf{x} = \mathbf{b}$. This problem breaks down into three possible outcomes: no solution, exactly one solution, and infinitely many solutions. The principal method for solving linear systems is row reduction. For small matrices, we can carry this out by hand; for large systems, computer software is required. Although we assume the reader is already familiar with the mechanics of row reduction, we review this topic briefly because we need to have a thorough understanding of the theory behind this method.

The last two sections of this chapter are devoted to determinants. Determinants have a mixed reputation in mathematics. On the one hand, they give us powerful tools for proving theorems in both geometry and linear algebra, and on the other hand they are costly to compute and thus give way to other more efficient methods for solving most linear algebra problems. One place where they have substantial value, however, is in understanding how linear transformations map volumes from one space to another. Indeed, determinants are used to define the Jacobians used in multidimensional integration; see Section 8.7.

In this chapter, we briefly study the classical theory of determinants and prove two major theorems—first, that the determinant of a matrix can be computed via row reduction to the simpler problem of computing a determinant of an upper-triangular matrix and, second, that the determinant of an upper-triangular matrix is just the product of its diagonal elements. These two results allow us to compute determinants fairly efficiently.

2.1 Basics of Linear Transformations I

Maps between vector spaces that preserve the vector-space structure are called *linear transformations*. Linear transformations are the main objects of study in linear algebra and are a key tool for understanding much of mathematics. Any linear transformation has two important subspaces associated with it, namely, the *kernel*, which consists of all the vectors in the domain that map to zero, and the *range* or image of the linear transformation. In this section we study the basic properties of a linear transformation and the associated kernel and range.

2.1.1 Definition and Examples

Definition 2.1.1. Let V and W be vector spaces over a common field $\mathbb{F}$. A map $L: V \rightarrow W$ is a linear transformation from V into W if

$$L(ax_1 + bx_2) = aL(x_1) + bL(x_2) \quad (2.1)$$

for all vectors $x_1, x_2 \in V$ and scalars $a, b \in \mathbb{F}$. We say that a linear transformation is a linear operator if it maps a vector space into itself.

Remark 2.1.2. As mentioned in the definition, linear transformations are always between vector spaces with a common scalar field. Maps between vector spaces with different scalar fields are rare. It is occasionally useful to think of the complex numbers as a two-dimensional real vector space, but in that case, we would state very explicitly which scalar field we are using, and even then the linear transformations we study would still have the same scalar field (whether $\mathbb{R}$ or $\mathbb{C}$) for both domain and codomain.

Example 2.1.3.

- (i) For any positive integer n , the projection map $\pi_i: \mathbb{F}^n \rightarrow \mathbb{F}$ given by $(a_1, \dots, a_n) \mapsto a_i$ is linear for each $i = 1, \dots, n$.
- (ii) For any positive integer n , the n -tuple $(a_1, \dots, a_n)$ of scalars defines a linear map $(x_1, \dots, x_n) \mapsto a_1x_1 + \dots + a_nx_n$ from $\mathbb{F}^n \rightarrow \mathbb{F}$.
- (iii) More generally, any $m \times n$ matrix A with entries in $\mathbb{F}$ defines a linear transformation $\mathbb{F}^n \rightarrow \mathbb{F}^m$ by the rule $\mathbf{x} \mapsto A\mathbf{x}$.
- (iv) For any positive integer n , the map $C^n((a, b); \mathbb{F}) \rightarrow C^{n-1}((a, b); \mathbb{F})$ defined by $f \mapsto \frac{d}{dx}f$ is linear.
- (v) The map $C([a, b]; \mathbb{F}) \rightarrow \mathbb{F}$ defined by $f \mapsto \int_a^b f(x) dx$ is linear.
- (vi) For any interval $[a, b] \subset \mathbb{R}$, the map $C([a, b]; \mathbb{F}) \rightarrow C([a, b]; \mathbb{F})$ given by $f \mapsto \int_a^x f(t) dt$ is linear.
- (vii) For any interval $[a, b] \subset \mathbb{R}$ and any $p \in [a, b]$, the evaluation map $e_p: C([a, b]; \mathbb{F}) \rightarrow \mathbb{F}$ defined by $f(x) \mapsto f(p)$ is linear.
- (viii) For any interval $[a, b] \subset \mathbb{R}$, a polynomial $p \in \mathbb{F}[x]$ is continuous on $[a, b]$, and thus the identity map $\mathbb{F}[x] \rightarrow C([a, b]; \mathbb{F})$ is linear.
- (ix) For any positive integer n , the projection map $\ell^p \rightarrow \mathbb{F}^n$ given by

$$(a_1, a_2, \dots, a_n, \dots) \mapsto (a_1, a_2, \dots, a_n)$$

is a linear transformation.

- (x) The left-shift map $\ell^p \rightarrow \ell^p$ given by

$$(a_1, a_2, \dots, a_n, \dots) \mapsto (a_2, a_3, \dots, a_n, \dots)$$

is a linear operator.

- (xi) For any vector space V defined over the field $\mathbb{F}$ and any scalar $\alpha \in \mathbb{F}$, the scaling operator $h_\alpha : V \rightarrow V$ given by $h_\alpha(\mathbf{x}) = \alpha\mathbf{x}$ is linear.
- (xii) For any $\theta \in [0, 2\pi)$, the map $\rho_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by rotating around the origin counterclockwise by angle θ is a linear operator. This can be written as

$$\rho_\theta(x, y) = (x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta).$$

- (xiii) For $A \in M_m(\mathbb{F})$ and $B \in M_n(\mathbb{F})$, the Sylvester map $L : M_{m \times n}(\mathbb{F}) \rightarrow M_{m \times n}(\mathbb{F})$ given by $X \mapsto AX + XB$ is a linear operator. Similarly, the Stein map $K : M_{m \times n}(\mathbb{F}) \rightarrow M_{m \times n}(\mathbb{F})$ given by $X \mapsto AXB - X$ is also a linear operator.

Unexample 2.1.4.

- (i) The map $\phi : \mathbb{F} \rightarrow \mathbb{F}$ given by $\phi(x) = x^2$ is not linear because $(x + y)^2$ is not equal to $x^2 + y^2$, and therefore $\phi(x + y) \neq \phi(x) + \phi(y)$.
- (ii) Most functions $\mathbb{F} \rightarrow \mathbb{F}$ are *not* linear operators; indeed, a continuous function $f : \mathbb{F} \rightarrow \mathbb{F}$ is linear if and only if $f(x) = ax$. Thus, most functions in the standard arsenal of useful functions, including $\cos(x)$, $\sin(x)$, x^n , e^x , and most other functions you can think of, are not linear. However, do not confuse this with the fact that the set $C(\mathbb{F}; \mathbb{F})$ of all continuous functions from $\mathbb{F}$ to $\mathbb{F}$ is a vector space, and there are many linear operators on this vector space. Such operators are not functions from $\mathbb{F}$ to $\mathbb{F}$, but rather are functions from $C(\mathbb{F}; \mathbb{F})$ to itself. In other words, the vectors in $C(\mathbb{F}; \mathbb{F})$ are continuous functions and the operators act on those continuous functions and return other continuous functions. For example, if $h \in C(\mathbb{F}; \mathbb{F})$ is any continuous function, we may define a linear operator $L_h : C(\mathbb{F}; \mathbb{F}) \rightarrow C(\mathbb{F}; \mathbb{F})$ by right multiplication; that is, $L_h[f] = h \cdot f$.

The fundamental properties of linear transformations are given in the following proposition.

Proposition 2.1.5. *Let V and W be vector spaces. A linear transformation $L : V \rightarrow W$*

- (i) *maps lines into lines; that is, $L(tx + (1 - t)y) = tL(\mathbf{x}) + (1 - t)L(\mathbf{y})$ for all $t \in \mathbb{F}$;*

- (ii) maps the origin to the origin; that is, $L(\mathbf{0}) = \mathbf{0}$; and
- (iii) maps subspaces to subspaces; that is, if X is a subspace of V , then the set $L(X)$ is a subspace of W .

Proof.

- (i) This is an immediate consequence of (2.1).
- (ii) Since $L(\mathbf{x}) = L(\mathbf{x} + \mathbf{0}) = L(\mathbf{x}) + L(\mathbf{0})$, it follows by the uniqueness of the additive identity that $L(\mathbf{0}) = \mathbf{0}$ by Proposition 1.1.7.
- (iii) Assume that $\mathbf{y}_1, \mathbf{y}_2 \in L(X)$. Hence, $\mathbf{y}_1 = L(\mathbf{x}_1)$ and $\mathbf{y}_2 = L(\mathbf{x}_2)$ for some $\mathbf{x}_1, \mathbf{x}_2 \in X$. Since $a\mathbf{x}_1 + b\mathbf{x}_2 \in X$, we have that $a\mathbf{y}_1 + b\mathbf{y}_2 = aL(\mathbf{x}_1) + bL(\mathbf{x}_2) = L(a\mathbf{x}_1 + b\mathbf{x}_2) \in L(X)$. $\square$

Remark 2.1.6. In Proposition 2.1.5(i), the line $t\mathbf{x} + (1-t)\mathbf{y}$ maps to the origin if $L(\mathbf{x}) = L(\mathbf{y}) = \mathbf{0}$. If we consider a point as a degenerate line, we can still say that linear transformations map lines to lines.

2.1.2 The Kernel and Range

A fundamental problem in linear algebra is to solve systems of linear equations. This problem can be recast as finding all solutions $\mathbf{x}$ to the equation $L(\mathbf{x}) = \mathbf{b}$, where L is a linear map from the vector space V into W and $\mathbf{b}$ is an element of W .

The special case when $\mathbf{b} = \mathbf{0}$ is called a *homogeneous* linear system. The solution set $\mathcal{N}(L) = \{\mathbf{x} \in V \mid L(\mathbf{x}) = \mathbf{0}\}$ is called the *kernel* of the transformation L . The system $L(\mathbf{x}) = \mathbf{0}$ always has at least one solution ($\mathbf{x} = \mathbf{0}$), and so $\mathcal{N}(L)$ is always nonempty. In fact, it turns out that $\mathcal{N}(L)$ is a subspace of the domain V .

On the other hand if $\mathbf{b} \neq \mathbf{0}$, it is possible that the system has no solutions at all. It all depends on whether $\mathbf{b}$ is in the *range* $\mathcal{R}(L)$ of the linear transformation L . As it turns out, the range also forms a subspace, but of W instead of V .

These two subspaces, the kernel and the range, tell us a lot about the linear transformation L and a lot about the solutions of linear systems of the form $L(\mathbf{x}) = \mathbf{b}$. For example, for each $\mathbf{b} \in \mathcal{R}(L)$ the set of all solutions of $L(\mathbf{x}) = \mathbf{b}$ is a coset $\mathbf{v} + \mathcal{N}(L)$ for some $\mathbf{v} \in V$.

Definition 2.1.7. Let V and W be vector spaces. The *kernel* (or *null space*) of a linear transformation $L : V \rightarrow W$ is the set $\mathcal{N}(L) = \{\mathbf{x} \in V \mid L(\mathbf{x}) = \mathbf{0}\}$. The *range* (or *image*) of L is the set $\mathcal{R}(L) = \{L(\mathbf{x}) \in W \mid \mathbf{x} \in V\}$.

Proposition 2.1.8. Let V and W be vector spaces, and let $L : V \rightarrow W$ be a linear map. We have the following:

- (i) $\mathcal{N}(L)$ is a subspace of V .
- (ii) $\mathcal{R}(L)$ is a subspace of W .

Proof. Note that both $\mathcal{N}(L)$ and $\mathcal{R}(L)$ are nonempty since $L(\mathbf{0}) = \mathbf{0}$.

- (i) If $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{N}(L)$, then $L(\mathbf{x}_1) = L(\mathbf{x}_2) = \mathbf{0}$, which implies that $L(a\mathbf{x}_1 + b\mathbf{x}_2) = aL(\mathbf{x}_1) + bL(\mathbf{x}_2) = \mathbf{0}$. Therefore $a\mathbf{x}_1 + b\mathbf{x}_2 \in \mathcal{N}(L)$.
- (ii) This follows immediately from Proposition 2.1.5(iii). $\square$

Example 2.1.9. We have the following:

- (i) Let $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be a projection given by $(x, y, z) \mapsto (x, y)$. It follows that $\mathcal{R}(L) = \mathbb{R}^2$ and $\mathcal{N}(L) = \{(0, 0, z) \mid z \in \mathbb{R}\}$.
- (ii) Let $L : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be an identity map given by $(x, y) \mapsto (x, y, 0)$. It follows that $\mathcal{R}(L) = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$ and $\mathcal{N}(L) = \{(0, 0)\}$.
- (iii) Let $L : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be given by $L(\mathbf{x}) = \mathbf{x}$. It follows that $\mathcal{N}(L) = \{\mathbf{0}\}$ and $\mathcal{R}(L) = \mathbb{R}^3$.
- (iv) Let $L : C^1([0, 1]; \mathbb{F}) \rightarrow C([0, 1]; \mathbb{F})$ be given by $L[f] = f' + f$. It is easy to show that L is linear. Note that $\mathcal{N}(L) = \text{span}\{e^{-x}\}$. To prove that L is surjective, let $g(x) \in C([0, 1]; \mathbb{F})$, and for $0 \leq x \leq 1$, define

$$f(x) = e^{-x} \int_0^x g(t)e^t dt + Ce^{-x}$$

for any $C \in \mathbb{F}$. It is straightforward to verify that $L[f] = g$. Thus, $\mathcal{R}(L) = C([0, 1]; \mathbb{F})$.

Remark 2.1.10. Given a linear map L from one function space to another, we typically write $L[f]$ as the image of f . To evaluate the function at a point x , we write $L[f](x)$.

2.2 Basics of Linear Transformations II

In this section, we examine two algebraic properties that allow us to build new linear transformations from existing ones: the composition of two linear maps is again a linear map, and a linear combination of linear maps is a linear map.

We also define formally what it means for two vector spaces to be “the same,” or *isomorphic*. Maps identifying isomorphic vector spaces are called *isomorphisms*. Isomorphisms are useful when one vector space is difficult or confusing to work with, but we can construct an isomorphism to another vector space where the structure is easier to visualize or work with. Any results we find in the new vector space also apply to the original vector space.

2.2.1 Two Algebraic Properties of Linear Transformations

Proposition 2.2.1.

- (i) If $L : V \rightarrow W$ and $K : W \rightarrow X$ are linear transformations, then the composition $K \circ L : V \rightarrow X$, is also a linear transformation.

- (ii) If $L : V \rightarrow W$ and $K : V \rightarrow W$ are linear transformations and $r, s \in \mathbb{F}$, then the map $rL + sK$, defined by $(rL + sK)(\mathbf{x}) = rL(\mathbf{x}) + sK(\mathbf{x})$, is also a linear transformation.

Proof.

- (i) If $a, b \in \mathbb{F}$ and $\mathbf{x}, \mathbf{y} \in V$, then

$$\begin{aligned}(K \circ L)(a\mathbf{x} + b\mathbf{y}) &= K(aL(\mathbf{x}) + bL(\mathbf{y})) = aK(L(\mathbf{x})) + bK(L(\mathbf{y})) \\ &= a(K \circ L)(\mathbf{x}) + b(K \circ L)(\mathbf{y}).\end{aligned}$$

Thus, the composition of two linear transformations is linear.

- (ii) If $a, b, r, s \in \mathbb{F}$ and $\mathbf{x}, \mathbf{y} \in V$, then

$$\begin{aligned}(rL + sK)(a\mathbf{x} + b\mathbf{y}) &= rL(a\mathbf{x} + b\mathbf{y}) + sK(a\mathbf{x} + b\mathbf{y}) \\ &= r(aL(\mathbf{x}) + bL(\mathbf{y})) + s(aK(\mathbf{x}) + bK(\mathbf{y})) \\ &= a(rL + sK)(\mathbf{x}) + b(rL + sK)(\mathbf{y}).\end{aligned}$$

Thus, the linear combination $rL + sK$ is itself a linear transformation. $\square$

Definition 2.2.2. Let V and W be vector spaces over the same field $\mathbb{F}$. Let $\mathcal{L}(V; W)$ be the set of linear transformations L mapping V into W with the pointwise operations of vector addition and scalar multiplication:

- (i) If $f, g \in \mathcal{L}(V; W)$, then $f + g$ is the map defined by $(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$.
(ii) If $a \in \mathbb{F}$, then af is the map defined by $(af)(\mathbf{x}) = af(\mathbf{x})$.

Corollary 2.2.3. Let V and W be vector spaces over the same field $\mathbb{F}$. The set $\mathcal{L}(V; W)$ with the pointwise operations of vector addition and scalar multiplication forms a vector space over $\mathbb{F}$.

Proof. The proof is Exercise 2.5. $\square$

Remark 2.2.4. For notational convenience, we denote the composition of two linear transformations K and L as the product KL instead of $K \circ L$. When expressing the repeated composition of linear operators $K : V \rightarrow V$ we write as powers of K ; for example, we write K^2 instead of $K \circ K$.

Nota Bene 2.2.5. The distributive laws hold for compositions of sums and sums of compositions; that is, $K(L_1 + L_2) = KL_1 + KL_2$ and $(L_1 + L_2)K = L_1K + L_2K$. However, commutativity generally fails. In fact, if $L : V \rightarrow W$ and $K : W \rightarrow X$ and $X \neq V$, then the composition LK doesn't even make sense. Even when $X = V$, we seldom have commutativity.

2.2.2 Invertibility

We finish this section by developing the main ideas of invertibility and defining the concept of an *isomorphism*, which tells us when two vector spaces are essentially the same.

Definition 2.2.6. A linear transformation $L : V \rightarrow W$ is called invertible if it has an inverse that is also a linear transformation.

A function has an inverse if and only if it is bijective; see Theorem A.2.19(iii) in the appendix. One might think that there could be a linear transformation whose inverse function is not linear, but the next proposition shows that if a linear transformation has an inverse, then that inverse must also be linear.

Proposition 2.2.7. If a linear transformation is bijective, then the inverse function is also a linear transformation.

Proof. Assume that $L : V \rightarrow W$ is a bijective linear transformation with inverse function L^{-1} . Given $\mathbf{w}_1, \mathbf{w}_2 \in W$ there exist $\mathbf{x}_1, \mathbf{x}_2 \in V$ such that $L(\mathbf{x}_1) = \mathbf{w}_1$ and $L(\mathbf{x}_2) = \mathbf{w}_2$. Thus, for $a, b \in \mathbb{F}$, we have $L^{-1}(a\mathbf{w}_1 + b\mathbf{w}_2) = L^{-1}(aL(\mathbf{x}_1) + bL(\mathbf{x}_2)) = L^{-1}(L(a\mathbf{x}_1 + b\mathbf{x}_2)) = a\mathbf{x}_1 + b\mathbf{x}_2 = aL^{-1}(\mathbf{w}_1) + bL^{-1}(\mathbf{w}_2)$. Thus, L^{-1} is linear. $\square$

Corollary 2.2.8. A linear transformation is invertible if and only if it is bijective.

Definition 2.2.9. An invertible linear transformation is called an isomorphism.⁸ Two spaces V and W are isomorphic, denoted $V \cong W$, if there exists an isomorphism $L : V \rightarrow W$. If an isomorphism is a linear operator (that is, if $W = V$), then it is called an automorphism.⁹

Example 2.2.10.

- (i) For any positive integer n , let $W = \{(0, a_2, a_3, \dots, a_n) \mid a_i \in \mathbb{F}\}$ be a subspace of $\mathbb{F}^n$. We claim $\mathbb{F}^{n-1}$ is isomorphic to W . The isomorphism between them is the linear map $(b_1, b_2, \dots, b_{n-1}) \mapsto (0, b_1, b_2, \dots, b_{n-1})$ with the obvious inverse.

⁸For experts, we note that it is common to define an isomorphism to be a bijective linear transformation, and although that is equivalent to our definition, it is the “wrong” definition. What we really care about is that all the properties of the two vector spaces match up—that is, any relation in one vector space maps to (by the isomorphism or by its inverse) the same relation in the other vector space. The importance of requiring the inverse to preserve all the important properties is evident in categories where bijectivity is not sufficient. For example, an isomorphism of topological spaces (a homeomorphism) is a continuous map that has a continuous inverse, but a continuous bijective map need not have a continuous inverse.

⁹It is worth noting the Greek roots of these words: *iso* means equal, *morph* means shape or form, and *auto* means self.

- (ii) The vector space $\mathbb{F}[x; n]$ of polynomials of degree at most n is isomorphic to $\mathbb{F}^{n+1}$. The isomorphism is given by the map $a_0 + a_1x + \cdots + a_nx^n \mapsto (a_0, a_1, \dots, a_n) \in \mathbb{F}^{n+1}$, which is readily seen to be linear and invertible.

Note that while these two vector spaces are isomorphic, the domain has additional structure that is not necessarily preserved by the vector-space isomorphism. For example, multiplication of polynomials is an interesting and useful operation, but it is not part of the vector-space structure, and so it is not guaranteed to be preserved by isomorphism.

Remark 2.2.11. The inverse of an invertible linear transformation L is also invertible. Specifically, $(L^{-1})^{-1} = L$. Also, the composition KL of two invertible linear transformations is itself invertible with inverse $(KL)^{-1} = L^{-1}K^{-1}$ (see Exercise 2.6).

2.2.3 Isomorphisms

Theorem 2.2.12. *The relation $\cong$ is an equivalence relation on the collection of all vector spaces.*

Proof. Since the identity map $I_V : V \rightarrow V$ is an isomorphism, it follows that $\cong$ is reflexive. If $V \cong W$, then there exists an isomorphism $L : V \rightarrow W$. By Remark 2.2.11, we know that L^{-1} is also an isomorphism, and thus $\cong$ is symmetric. Finally, if $V \cong W$ and $W \cong X$, then there exist isomorphisms $L : V \rightarrow W$ and $K : W \rightarrow X$. By Remark 2.2.11 the composition of isomorphisms is an isomorphism, which shows that KL is an isomorphism and $\cong$ is transitive. $\square$

Remark 2.2.13. Of course not every operator is invertible, but even when an operator is not invertible, many of the results of invertibility can still be used. In Sections 4.6.1 and 12.9, we examine certain *generalized inverses* of linear operators. These are operators that are not quite inverses but do have some of the properties enjoyed by inverses.

Two vector spaces that are isomorphic are essentially the same with respect to the linear structure. The following shows that every property of vector spaces that we have discussed so far in this book is preserved by isomorphism. We use this often because many vector spaces that look complicated can be shown to be isomorphic to a simpler-looking vector space. In particular, we show in Corollary 2.3.12 that all vector spaces of dimension n are isomorphic.

Proposition 2.2.14. *If V and W are isomorphic vector spaces, with isomorphism $L : V \rightarrow W$, then the following hold:*

- (i) *A linear equation holds in V if and only if it also holds in W ; that is, $\sum_{i=1}^{\ell} a_i \mathbf{x}_i = \mathbf{0}$ holds in V if and only if $\sum_{i=1}^{\ell} a_i L(\mathbf{x}_i) = \mathbf{0}$ holds in W .*
- (ii) *A set $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis of V if and only if $LB = \{L(\mathbf{v}_1), \dots, L(\mathbf{v}_n)\}$ is a basis for W . Moreover, the dimension of V is equal to the dimension of W .*

- (iii) The set of all subspaces of V is in bijective correspondence with the set of all subspaces of W .
- (iv) If $K : W \rightarrow X$ is any linear transformation, then the composition $KL : V \rightarrow X$ is also a linear transformation, and we have

$$\mathcal{N}(KL) = L^{-1}\mathcal{N}(K) = \{\mathbf{v} \mid L(\mathbf{v}) \in \mathcal{N}(K)\}$$

and

$$\mathcal{R}(KL) = \mathcal{R}(K).$$

Proof.

- (i) If $\sum_{i=1}^{\ell} a_i \mathbf{x}_i = \mathbf{0}$ in V , then $\mathbf{0} = L(\mathbf{0}) = L(\sum_{i=1}^{\ell} a_i \mathbf{x}_i) = \sum_{i=1}^{\ell} a_i L(\mathbf{x}_i)$ in W . The converse follows by applying L^{-1} to $\sum_{i=1}^{\ell} a_i L(\mathbf{x}_i) = \mathbf{0}$.
- (ii) Since L is surjective, any element $\mathbf{w} \in W$ can be written as $L(\mathbf{v})$ for some $\mathbf{v} \in V$. Because B is a basis, it spans V , and we can write $\mathbf{v} = \sum_{i=1}^n a_i \mathbf{v}_i$ for some choice of $a_1, \dots, a_n \in \mathbb{F}$. Applying L gives $\mathbf{w} = L(\mathbf{v}) = L(\sum_{i=1}^n a_i \mathbf{v}_i) = \sum_{i=1}^n a_i L(\mathbf{v}_i)$, so the set $LB = \{L(\mathbf{v}_1), \dots, L(\mathbf{v}_n)\}$ spans W . It is also linearly independent since $\sum_{i=1}^n c_i L(\mathbf{v}_i) = \mathbf{0}$ implies $\sum_{i=1}^n c_i L^{-1}L(\mathbf{v}_i) = \mathbf{0}$ by part (i). Since B is linearly independent, we must have $c_i = 0$ for all i .

The converse follows by applying the same argument with L^{-1} .

- (iii) Let $\mathcal{S}_V$ be the set of subspaces of V and $\mathcal{S}_W$ be the set of subspaces of W . The linear transformation L induces a map $\tilde{L} : \mathcal{S}_V \rightarrow \mathcal{S}_W$ given by sending $X \in \mathcal{S}_V$ to the subspace $LX = \{L(\mathbf{v}) \mid \mathbf{v} \in X\} \in \mathcal{S}_W$. Similarly, L^{-1} induces a map $\tilde{L}^{-1} : \mathcal{S}_W \rightarrow \mathcal{S}_V$. The composition $\tilde{L}^{-1}\tilde{L}$ is the identity $I_{\mathcal{S}_V}$ because for any $X \in \mathcal{S}_V$ we have

$$\tilde{L}^{-1}\tilde{L}(X) = \tilde{L}^{-1}\{L(\mathbf{v}) \mid \mathbf{v} \in X\} = \{L^{-1}L\mathbf{v} \mid \mathbf{v} \in X\} = \{\mathbf{v} \mid \mathbf{v} \in X\} = X.$$

Similarly, $\tilde{L}\tilde{L}^{-1} = I_{\mathcal{S}_W}$, so the maps $\tilde{L}$ and $\tilde{L}^{-1}$ are inverses and thus bijective.

- (iv) See Exercise 2.7. $\square$

2.3 Rank, Nullity, and the First Isomorphism Theorem

As we saw in the previous section, the kernel is the solution set to the *homogeneous linear system* $L(\mathbf{x}) = \mathbf{0}$, and it forms a subspace of the domain of L . When we want to consider the more general equation $L(\mathbf{x}) = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$, the solution set is no longer a subspace, but rather a coset of the kernel. That is, if $\mathbf{x}_0$ is any one solution to the system $L(\mathbf{x}) = \mathbf{b}$, then the set of all possible solutions is the coset $\mathbf{x}_0 + \mathcal{N}(L)$. This illustrates the importance of the quotient space $V/\mathcal{N}(L)$ in the study of linear transformations and linear systems of equations.

In this section, we study these relationships more carefully. In particular, we show that the quotient $V/\mathcal{N}(L)$ is isomorphic to $\mathcal{R}(L)$. This is called the *first isomorphism theorem*, and it has powerful consequences. We use it to give a

formula relating the dimension of the image (called the *rank*) and the dimension of the kernel (called the *nullity*).

We also use the first isomorphism theorem to prove the extremely important result that all n -dimensional vector spaces over $\mathbb{F}$ are isomorphic to $\mathbb{F}^n$.

Finally, we use the first isomorphism theorem to prove another theorem, called the *second isomorphism theorem*, about the relation between quotients of sums and intersections of subspaces. This theorem provides a formula, called the *dimension formula*, relating the dimensions of sums and intersections of subspaces.

2.3.1 Quotients, Rank, and Nullity

Proposition 2.3.1. *Let W be a subspace of the vector space V and denote by V/W the quotient space. The mapping $\pi : V \rightarrow V/W$ defined by $\pi(\mathbf{v}) = \mathbf{v} + W$ is a surjective linear transformation. We call π the canonical epimorphism.¹⁰*

Proof. Since π is clearly surjective, it suffices to show that π is linear. We check $\pi(a\mathbf{x} + b\mathbf{y}) = (a\mathbf{x} + b\mathbf{y}) + W = a \square (\mathbf{x} + W) \boxplus b \square (\mathbf{y} + W) = a \square \pi(\mathbf{x}) \boxplus b \square \pi(\mathbf{y})$. $\square$

Example 2.3.2. Exercise 1.29 shows for a vector space V that $V/\{\mathbf{0}\} \cong V$. Alternatively, this follows from the proposition as follows. Define a linear map $\pi : V \rightarrow V/\{\mathbf{0}\}$ by $\pi(\mathbf{x}) = \mathbf{x} + \{\mathbf{0}\}$. By the proposition, π is surjective. Thus, it suffices to show that π is injective. If $\pi(\mathbf{x}) = \pi(\mathbf{y})$, then $\mathbf{x} + \{\mathbf{0}\} = \mathbf{y} + \{\mathbf{0}\}$, which implies that $\mathbf{x} - \mathbf{y} \in \{\mathbf{0}\}$. Therefore, $\mathbf{x} = \mathbf{y}$ and π is injective.

The following lemma is a very handy tool.

Lemma 2.3.3. *A linear map L is injective if and only if $\mathcal{N}(L) = \{\mathbf{0}\}$.*

Proof. L is injective if and only if $L(\mathbf{x}) = L(\mathbf{y})$ implies $\mathbf{x} = \mathbf{y}$, which holds if and only if $L(\mathbf{z}) = 0$ implies $\mathbf{z} = 0$, which holds if and only if $\mathcal{N}(L) = \{\mathbf{0}\}$. $\square$

Lemma 2.3.4. *Let $L : V \rightarrow Z$ be a linear transformation between vector spaces V and Z . Assume that W and Y are subspaces of V and Z , respectively. If $L(W) \subset Y$, then L induces a linear transformation $\bar{L} : V/W \rightarrow Z/Y$ defined by $\bar{L}(\mathbf{x} + W) = L(\mathbf{x}) + Y$.*

Proof. We show that $\bar{L}$ is well defined (see Appendix A.2.4) and linear. If $\mathbf{x} + W = \mathbf{y} + W$, then $\mathbf{x} - \mathbf{y} \in W$, which implies that $L(\mathbf{x}) - L(\mathbf{y}) \in L(W) \subset Y$. Hence, $\bar{L}(\mathbf{x} + W) = L(\mathbf{x}) + Y = L(\mathbf{y}) + Y = \bar{L}(\mathbf{y} + W)$. Thus, $\bar{L}$ is well defined. To show

¹⁰An epimorphism is a surjective linear transformation. The name comes from the Greek root *epi-*, meaning *on*, *upon*, or *above*.

that $\bar{L}$ is linear, we note that

$$\begin{aligned}
 \bar{L}(a \square (\mathbf{x} + W) \boxplus b \square (\mathbf{y} + W)) &= \bar{L}(a\mathbf{x} + b\mathbf{y} + W) \\
 &= L(a\mathbf{x} + b\mathbf{y}) + Y \\
 &= aL(\mathbf{x}) + bL(\mathbf{y}) + Y \\
 &= a \square (L(\mathbf{x}) + Y) \boxplus b \square (L(\mathbf{y}) + Y) \\
 &= a \square \bar{L}(\mathbf{x} + W) \boxplus b \square \bar{L}(\mathbf{y} + W). \quad \square
 \end{aligned}$$

Theorem 2.3.5 (First Isomorphism Theorem). *If V and X are vector spaces and $L : V \rightarrow X$ is a linear transformation, then $V/\mathcal{N}(L) \cong \mathcal{R}(L)$. In particular, if L is surjective, then $V/\mathcal{N}(L) \cong X$.*

Proof. Apply the previous lemma with $W = \mathcal{N}(L)$, with $Y = \{\mathbf{0}\}$, and with $Z = \mathcal{R}(L)$ to get an induced linear transformation $\bar{L} : V/\mathcal{N}(L) \rightarrow \mathcal{R}(L)/\{\mathbf{0}\} \cong \mathcal{R}(L)$ that is clearly surjective since $\mathbf{x} + \mathcal{N}(L)$ maps to $L(\mathbf{x}) + \{\mathbf{0}\}$, which then maps to $L(\mathbf{x}) \in \mathcal{R}(L)$. Thus, it suffices to show that $\bar{L}$ is injective. If $\bar{L}(\mathbf{x} + \mathcal{N}(L)) = \bar{L}(\mathbf{y} + \mathcal{N}(L))$, then $L(\mathbf{x}) + \{\mathbf{0}\} = L(\mathbf{y}) + \{\mathbf{0}\}$. Equivalently, $L(\mathbf{x} - \mathbf{y}) \in \{\mathbf{0}\}$, which implies that $\mathbf{x} - \mathbf{y} \in \mathcal{N}(L)$, and so $\mathbf{x} + \mathcal{N}(L) = \mathbf{y} + \mathcal{N}(L)$. Thus, $\bar{L}$ is injective. $\square$

The above theorem is incredibly useful because we can use it to prove that a quotient V/W is isomorphic to another space X . The standard way to do this is to construct a surjective linear transformation $V \rightarrow X$ that has kernel equal to W . The first isomorphism theorem then gives the desired isomorphism.

Example 2.3.6.

- (i) If V is a vector space, then we can show that $V/V \cong \{\mathbf{0}\}$. Define the linear map $L : V \rightarrow \{\mathbf{0}\}$ by $L(\mathbf{x}) = \mathbf{0}$. The kernel of L is all of V , and so the result follows from the first isomorphism theorem.
- (ii) Let $V = \{(0, a_2, a_3, \dots, a_n) \mid a_i \in \mathbb{F}\} \subset \mathbb{F}^n$ for some integer $n \geq 2$. The quotient $\mathbb{F}^n/V$ is isomorphic to $\mathbb{F}$. To see this use the map $\pi_1 : \mathbb{F}^n \rightarrow \mathbb{F}$ (see Example 2.1.3(i)) defined by $(a_1, \dots, a_n) \mapsto a_1$. One can readily check that this is linear and surjective with kernel equal to V , and so the first isomorphism theorem gives the desired result.
- (iii) The set of constant functions $\text{Const}_{\mathbb{F}}$ forms a subspace of $C^n([a, b]; \mathbb{F})$. The quotient $C^n([a, b]; \mathbb{F})/\text{Const}_{\mathbb{F}}$ is isomorphic to $C^{n-1}([a, b]; \mathbb{F})$. To see this, define the linear map $D : C^n([a, b]; \mathbb{F}) \rightarrow C^{n-1}([a, b]; \mathbb{F})$ as the derivative $D[f](x) = f'(x)$. Its kernel is precisely $\text{Const}_{\mathbb{F}}$, and so the first isomorphism theorem implies the quotient is isomorphic to $\mathcal{R}(D)$. But D is also surjective, since, by the fundamental theorem of calculus, it is a right inverse to $\text{Int} : f \mapsto \int_a^x f(t) dt$ (see Example A.2.22). Since D is surjective, the induced map $\bar{D} : C^n([a, b]; \mathbb{F})/\text{Const}_{\mathbb{F}} \rightarrow C^{n-1}([a, b]; \mathbb{F})$ is an isomorphism.

- (iv) For any $p \in [a, b] \subset \mathbb{R}$ let $M_p = \{f \in C([a, b]; \mathbb{F}) \mid f(p) = 0\}$. We claim that $C([a, b]; \mathbb{F})/M_p \cong \mathbb{F}$. Use the linear map $e_p : C([a, b]; \mathbb{F}) \rightarrow \mathbb{F}$ given by $e_p(f) = f(p)$. The kernel of e_p is precisely M_p , and so the result follows from the first isomorphism theorem.

Theorem 2.3.7. *If W is a subspace of a finite-dimensional vector space V , then*

$$\dim V = \dim W + \dim V/W. \quad (2.2)$$

Proof. If $W = \{0\}$ or $W = V$, then the result follows trivially from Example 2.3.2 or Example 2.3.6(i). Thus, we assume that W is a proper, nontrivial subspace of V . Let $S = \{\mathbf{x}_1, \dots, \mathbf{x}_r\}$ be a basis for W . By the extension theorem (Corollary 1.4.5) we can choose $T = \{\mathbf{y}_{r+1}, \dots, \mathbf{y}_n\}$ so that $S \cup T$ is a basis for V and $S \cap T = \emptyset$. We claim that the set $T_W = \{\mathbf{y}_{r+1} + W, \dots, \mathbf{y}_n + W\}$ is a basis for V/W . It follows that $\dim V = n = r + (n - r) = \dim W + \dim V/W$.

We show that T_W is a basis. To show that T_W is linearly independent, assume

$$\beta_{r+1} \sqcap (\mathbf{y}_{r+1} + W) \boxplus \dots \boxplus \beta_n \sqcap (\mathbf{y}_n + W) = \mathbf{0} + W.$$

Thus, $\sum_{j=r+1}^n \beta_j \mathbf{y}_j \in W$, which implies that $\sum_{j=r+1}^n \beta_j \mathbf{y}_j = \mathbf{0}$ (since S spans W , not T). However since T is linearly independent, we have that each $\beta_j = 0$. Thus, T_W is linearly independent. To show that T_W spans V/W , let $\mathbf{v} + W \in V/W$ for some $\mathbf{v} \in V$. Thus, we have that $\mathbf{v} = \sum_{i=1}^r \alpha_i \mathbf{x}_i + \sum_{j=r+1}^n \beta_j \mathbf{y}_j$. Hence, $\mathbf{v} + W = \beta_{r+1} \sqcap (\mathbf{y}_{r+1} + W) \boxplus \dots \boxplus \beta_n \sqcap (\mathbf{y}_n + W)$. $\square$

Definition 2.3.8. *Let $L : V \rightarrow W$ be a linear transformation between vector spaces V and W . The rank of L is the dimension of its range, and the nullity of L is the dimension of its kernel. More precisely, $\text{rank}(L) = \dim \mathcal{R}(L)$ and $\text{nullity}(L) = \dim \mathcal{N}(L)$.*

Corollary 2.3.9 (Rank-Nullity Theorem). *Let V and W be vector spaces, with V of finite dimension. If $L : V \rightarrow W$ is a linear transformation, then*

$$\dim V = \dim \mathcal{R}(L) + \dim \mathcal{N}(L) = \text{rank}(L) + \text{nullity}(L). \quad (2.3)$$

Proof. Note that (2.2) implies that $\dim V = \dim \mathcal{N}(L) + \dim V/\mathcal{N}(L)$. The first isomorphism theorem (Theorem 2.3.5) tells us that $V/\mathcal{N}(L) \cong \mathcal{R}(L)$, and dimension is preserved by isomorphism (Proposition 2.2.14(ii)), so $\dim V/\mathcal{N}(L) = \dim \mathcal{R}(L)$. $\square$

Remark 2.3.10. So far, when talking about addition and scalar multiplication of cosets we have used the notation $\boxplus$ and $\sqcap$ in order to help the reader see that these operations on V/W differ from the addition and scalar multiplication in V . However, most authors use $+$ and $\cdot$ (or juxtaposition) for these operations on the quotient space and just expect the reader to be able to identify from context when the operators are being used in V/W and when they are being used

in V . From now on we also use this more standard, but possibly more confusing, notation.

2.3.2 Isomorphisms of Finite-Dimensional Vector Spaces

Corollary 2.3.11. *Assume that V and W are n -dimensional vector spaces. A linear map $L : V \rightarrow W$ is injective if and only if it is surjective.*

Proof. By Lemma 2.3.3, L is injective if and only if $\dim(\mathcal{N}(L)) = 0$. By the rank-nullity theorem, this holds if and only if $\text{rank}(L) = n$. However, by Corollary 1.4.7 we know $\text{rank}(L) = n$ if and only if $\mathcal{R}(L) = W$, which means L is surjective. $\square$

Corollary 2.3.11 implies that two vector spaces of the same finite dimension, and over the same field $\mathbb{F}$, are isomorphic if there is either an injective or a surjective linear mapping between the two spaces. The next corollary shows that such a map always exists.

Corollary 2.3.12. *An n -dimensional vector space V over the field $\mathbb{F}$ is isomorphic to $\mathbb{F}^n$.*

Proof. Let $T = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a basis for V . Define the map $L : \mathbb{F}^n \rightarrow V$ as $L((a_1, a_2, \dots, a_n)) = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n$. It is straightforward to check that this is a linear transformation. We wish to show that L is bijective and hence an isomorphism. By Corollary 2.3.11, it suffices to show that L is surjective. Because T is a basis, every element $\mathbf{x} \in V$ can be written as a linear combination $\mathbf{x} = \sum_{i=1}^n a_i\mathbf{x}_i$ of vectors in T , so $\mathbf{x} = L(a_1, \dots, a_n)$. Hence, L is surjective, as required. $\square$

Example 2.3.13. An immediate consequence of the corollary above is that $\mathbb{F}[x; n] \cong \mathbb{F}^{n+1}$ and $M_{m \times n}(\mathbb{F}) \cong \mathbb{F}^{mn}$.

Remark 2.3.14. Corollary 2.3.12 is a big deal. Among other things, it means that even though there are many different *descriptions* of finite-dimensional vector spaces, there is essentially (up to isomorphism) only *one* n -dimensional vector space over $\mathbb{F}$ for each nonnegative integer n . When combined with the results of the next two sections, this allows essentially all of finite-dimensional linear algebra to be reduced to matrix analysis.

Remark 2.3.15. Again, we note that many vector spaces also carry additional structure that is not part of the vector-space structure, and the isomorphisms we have discussed do not necessarily preserve this other structure. So, for example, while $\mathbb{F}[x; n^2 - 1]$ is isomorphic to $\mathbb{F}^{n^2}$ as a vector space, and these are both isomorphic to $M_n(\mathbb{F})$ as vector spaces, multiplication of polynomials and multiplication of matrices in these vector spaces are not “the same.” In fact, multiplication of polynomials is commutative and multiplication of matrices is not, so we can’t hope to identify these multiplicative structures. But when we only care about the vector-space structure, the spaces are isomorphic and can be treated as being “the same.”

Nota Bene 2.3.16. Although any n -dimensional vector space is isomorphic to $\mathbb{F}^n$, beware that the actual isomorphism depends on the choice of basis and, in fact, on the choice of the *order* of the elements in the basis. If we had chosen a different basis or a different ordering of the elements of the basis in the proof of Corollary 2.3.12, we would have had a different isomorphism.

As a final consequence of Corollary 2.3.11, we show that for finite-dimensional vector spaces, if any operator has a left inverse, then that left inverse is also a right inverse.

Proposition 2.3.17. *Let $K, L : V \rightarrow V$ be linear operators on a finite-dimensional vector space V . If $KL = I$ is the identity map, then LK is also the identity map, and thus L and K are isomorphisms.*

Proof. The identity map is injective, so if $KL = I$, then L is also injective (see Proposition A.2.9), and by Corollary 2.3.11, the map L must also be surjective. Consider now the operator $(I - LK) : V \rightarrow V$. We have

$$(I - LK)L = L - L(KL) = L - L = 0.$$

This implies that every element of $\mathcal{R}(L)$ is in the kernel of $(I - LK)$. But L is surjective, so $\mathcal{N}(I - LK) = \mathcal{R}(L) = V$. Hence, $(I - LK) = 0$, and $I = LK$. $\square$

2.3.3 The Dimension Formula*

Corollary 2.3.18 (Second Isomorphism Theorem). *Let V be a vector space. If V_1 and V_2 are subspaces of V , then*

$$V_1/(V_1 \cap V_2) \cong (V_1 + V_2)/V_2. \quad (2.4)$$

The following diagram is sometimes helpful when thinking about the second isomorphism theorem:

$$\begin{array}{ccc} V_1 & \hookrightarrow & V_1 + V_2 \\ \uparrow & & \uparrow \\ V_1 \cap V_2 & \hookrightarrow & V_2 \end{array}$$

Here each of the symbols $\hookrightarrow$ denotes the inclusion of a subspace inside of another vector space. Associated with each inclusion $A \hookrightarrow B$ we can construct the quotient vector space B/A . The second isomorphism theorem says that the quotient associated with the left vertical side of the square is isomorphic to the quotient associated with the right vertical side of the square. If we relabel the subspaces, the second isomorphism theorem also says that the bottom quotient is isomorphic to the top quotient.

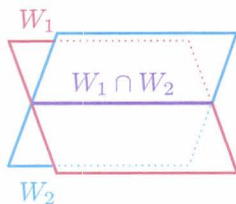


Figure 2.1. If a subspace W_1 has dimension d_1 and a subspace W_2 has dimension d_2 , then the dimension formula (2.5) says that the intersection $W_1 \cap W_2$ has dimension $\dim W_1 + \dim W_2 - \dim(W_1 + W_2)$. In this example, both W_1 and W_2 have dimension 2, and their sum has dimension 3; therefore the intersection has dimension 1.

Proof. Define $L : V_1 \rightarrow (V_1 + V_2)/V_2$ by $L(\mathbf{x}) = \mathbf{x} + V_2$. Clearly, L is surjective. Note that $\mathcal{N}(L) = \{\mathbf{x} \in V_1 \mid \mathbf{x} + V_2 = \mathbf{0} + V_2\} = \{\mathbf{x} \in V_1 \mid \mathbf{x} \in V_2\} = V_1 \cap V_2$. Thus, by Theorem 2.3.5, we have (2.4). $\square$

The next corollary tells us about the dimension of sums and intersections of subspaces. This result should seem geometrically intuitive: for example, if W_1 and W_2 are two distinct planes (through the origin) in $\mathbb{R}^3$, then their sum is all of $\mathbb{R}^3$, and their intersection is a line. Thus, we have $\dim W_1 + \dim W_2 = 2 + 2 = 4$, and $\dim(W_1 \cap W_2) + \dim(W_1 + W_2) = 1 + 3 = 4$; see Figure 2.1.

Corollary 2.3.19 (Dimension Formula). If V_1 and V_2 are finite-dimensional subspaces of a vector space V , then

$$\dim V_1 + \dim V_2 = \dim(V_1 \cap V_2) + \dim(V_1 + V_2). \quad (2.5)$$

Proof. By the second isomorphism theorem (Corollary 2.3.18), it follows that $\dim((V_1 + V_2)/V_2) = \dim(V_1/(V_1 \cap V_2))$. Also, by Theorem 2.3.7, it follows that $\dim V_1 = \dim(V_1 \cap V_2) + \dim(V_1/(V_1 \cap V_2))$. Therefore,

$$\begin{aligned} \dim V_1 + \dim V_2 &= \dim(V_1 \cap V_2) + \dim(V_1/(V_1 \cap V_2)) + \dim V_2 \\ &= \dim(V_1 \cap V_2) + \dim((V_1 + V_2)/V_2) + \dim V_2 \\ &= \dim(V_1 \cap V_2) + \dim(V_1 + V_2), \end{aligned}$$

where the last line again follows from Theorem 2.3.7. $\square$

2.4 Matrix Representations

In this section, we show how to represent linear transformations between finite-dimensional vector spaces as matrices. Given a finite ordered basis for each of the domain and codomain, a linear transformation has a unique matrix representation. In this representation, matrix-vector multiplication describes how the coordinates are mapped, and matrix-matrix multiplication corresponds to composition of linear

transformations. We first define transition matrices, which correspond to the matrix representation of the identity operator, but with two different bases for the domain and codomain. We then extend the construction to general linear transformations.

2.4.1 Transition Matrices

We begin by introducing some notation. Suppose that $S = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m]$ and $T = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m]$ are both ordered¹¹ bases of the vector space V . Any $\mathbf{x} \in V$ can be uniquely represented in each basis:

$$\mathbf{x} = \sum_{i=1}^m a_i \mathbf{s}_i \quad \text{and} \quad \mathbf{x} = \sum_{j=1}^m b_j \mathbf{t}_j.$$

In matrix notation, we may write this as

$$\mathbf{x} = [\mathbf{s}_1, \dots, \mathbf{s}_m] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \sum_{i=1}^m a_i \mathbf{s}_i \quad \text{and} \quad \mathbf{x} = [\mathbf{t}_1, \dots, \mathbf{t}_m] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \sum_{i=1}^m b_i \mathbf{t}_i.$$

We call the a_i above the *coordinates* of $\mathbf{x}$ in the basis S and denote the $m \times 1$ (column) matrix of these coordinates by $[\mathbf{x}]_S$. We proceed similarly for T :

$$[\mathbf{x}]_S = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \quad \text{and} \quad [\mathbf{x}]_T = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Since S is a basis, then by Corollary 1.2.15, each element of T can be uniquely expressed as a linear combination of elements of S ; we denote this as $\mathbf{t}_j = \sum_{i=1}^m c_{ij} \mathbf{s}_i$. In matrix notation, we write this as

$$[\mathbf{t}_1, \dots, \mathbf{t}_m] = [\mathbf{s}_1, \dots, \mathbf{s}_m] \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mm} \end{bmatrix}.$$

Thus,

$$\mathbf{x} = \sum_{j=1}^m b_j \mathbf{t}_j = \sum_{j=1}^m b_j \sum_{i=1}^m c_{ij} \mathbf{s}_i = \sum_{i=1}^m \left(\sum_{j=1}^m c_{ij} b_j \right) \mathbf{s}_i = \sum_{i=1}^m a_i \mathbf{s}_i.$$

¹¹In what follows, we want to keep track of the order of the elements of a set. Strictly speaking, a set isn't an object that maintains order—for example, the set $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is equal to the set $\{\mathbf{z}, \mathbf{y}, \mathbf{x}\}$ —and so we use square brackets to mean that the set is ordered. Thus, $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ means that $\mathbf{x}$ is the first element, $\mathbf{y}$ is the second element, and so on. That said, we do not always put the word “ordered” in front of the word basis—it is implied by the use of square brackets.

In terms of coordinates $[\mathbf{x}]_S$ and $[\mathbf{x}]_T$, this takes the form

$$[\mathbf{x}]_S = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = C[\mathbf{x}]_T,$$

where C is the matrix $[c_{ij}]$. Hence, C transforms the coordinates from the basis T into coordinates in the basis S . We call C the *transition matrix* from T into S and sometimes denote it with the subscripts C_{ST} to indicate that it provides the S -coordinates for a vector written originally in terms of T , that is,

$$[\mathbf{x}]_S = C_{ST}[\mathbf{x}]_T.$$

It is straightforward to verify that the transition matrix C_{TS} from S into T is given by the matrix inverse $C_{TS} = (C_{ST})^{-1}$; see Appendix C.1.3. This allows us to go back and forth between coordinates in S and in T , as demonstrated in the following example.

Example 2.4.1. Assume that $S = [x^2 - 1, x + 1, x - 1]$ and $T = [x^2, x, 1]$ are bases for $\mathbb{F}[x; 2]$. If $q(x) = 2(x^2 - 1) + 3(x + 1) - 5(x - 1)$, we can write q in the basis S as $[q]_S = [2 \ 3 \ -5]^T$. The transition matrix from S into T satisfies

$$[x^2 - 1, x + 1, x - 1] = [x^2, x, 1] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix}.$$

Thus, we can rewrite q in the basis T by matrix multiplication, and we find

$$[q]_T = C_{TS}[q]_S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ -5 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 6 \end{bmatrix}.$$

This corresponds to the equality $q(x) = 2(x^2 - 1) + 3(x + 1) - 5(x - 1) = 2x^2 - 2x + 6$. We can invert C_{TS} to get

$$C_{ST} = (C_{TS})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

which implies that

$$[x^2, x, 1] = [x^2 - 1, x + 1, x - 1] \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

This gives

$$[q]_S = C_{ST}[q]_T = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ -2 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -5 \end{bmatrix},$$

which corresponds to $q(x) = 2x^2 - 2x + 6 = 2(x^2 - 1) + 3(x + 1) - 5(x - 1)$.

Example 2.4.2. If $S = T$, then the matrix C_{SS} is just the identity matrix.

Example 2.4.3. To illustrate the importance of order, assume that the vectors of S and T are the same but are ordered differently. For example, if $m = 2$ and if $T = [s_2, s_1]$ is obtained from $S = [s_1, s_2]$ by switching the basis vectors, then the transition matrix is $C_{ST} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

2.4.2 Constructing the Matrix Representation

The next theorem tells us that a given linear map between two finite-dimensional vector spaces with ordered bases for the domain and codomain has a unique matrix representation.

Theorem 2.4.4. *Let V and W be finite-dimensional vector spaces over the field $\mathbb{F}$ with bases $S = [s_1, s_2, \dots, s_m]$ and $T = [t_1, t_2, \dots, t_n]$, respectively. Given a linear transformation $L : V \rightarrow W$, there exists a unique $n \times m$ matrix M_{TS} describing L in terms of the bases S and T ; that is, there exists a unique matrix M_{TS} such that*

$$[L(\mathbf{x})]_T = M_{TS}[\mathbf{x}]_S$$

for all $\mathbf{x} \in V$. We say that M_{TS} is the matrix representation of L from S into T .

Proof. If $\mathbf{x} = \sum_{j=1}^m a_j s_j$, then $L(\mathbf{x}) = \sum_{j=1}^m a_j L(s_j)$. Since each $L(s_j)$ can be written uniquely as a linear combination of elements of T , we have that $L(s_j) = \sum_{i=1}^n c_{ij} t_i$ for some matrix $M = [c_{ij}]$. Thus,

$$L(\mathbf{x}) = \sum_{j=1}^m a_j \left(\sum_{i=1}^n c_{ij} t_i \right) = \sum_{i=1}^n \left(\sum_{j=1}^m c_{ij} a_j \right) t_i = \sum_{i=1}^n b_i t_i,$$

where $b_i = \sum_{j=1}^m c_{ij} a_j$. In matrix notation, we have

$$[L(\mathbf{x})]_T = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = M[\mathbf{x}]_S.$$

To show uniqueness, we suppose that two matrix representations M and $\widehat{M}$ of L exist; that is, both $[L(\mathbf{x})]_T = M[\mathbf{x}]_S$ and $[L(\mathbf{x})]_T = \widehat{M}[\mathbf{x}]_S$ hold for all $\mathbf{x}$. This gives $(M - \widehat{M})[\mathbf{x}]_S = \mathbf{0}$ for all $\mathbf{x}$. However, this can only happen if $M - \widehat{M} = \mathbf{0}$. $\square$

Nota Bene 2.4.5. It is very important to remember that the matrix representation of L depends on the choices of *ordered* bases S and T . Different bases and different orderings almost always give different matrix representations.

Remark 2.4.6. The transition matrix we defined in the previous section is just the matrix representation of the identity transformation I_V , but with basis S used to represent the inputs of I_V and basis T used to represent the outputs of I_V .

Example 2.4.7. Let $S = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3]$ and $T = [\mathbf{t}_1, \mathbf{t}_2]$ be bases for $\mathbb{F}^3$ and $\mathbb{F}^2$, respectively. Consider the linear map $L : \mathbb{F}^3 \rightarrow \mathbb{F}^2$ that satisfies

$$[L(\mathbf{s}_1)]_T = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad [L(\mathbf{s}_2)]_T = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad \text{and} \quad [L(\mathbf{s}_3)]_T = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}.$$

We can write

$$[L(\mathbf{s}_1), L(\mathbf{s}_2), L(\mathbf{s}_3)] = [\mathbf{t}_1, \mathbf{t}_2]C_{TS} = [\mathbf{t}_1, \mathbf{t}_2] \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix},$$

and C_{TS} is the matrix representation of L from S into T . We compute $L(\mathbf{x})$ for $\mathbf{x} = 3\mathbf{s}_1 + 2\mathbf{s}_2 + 4\mathbf{s}_3$. Note that

$$C_{TS}[\mathbf{x}]_S = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} = [L(\mathbf{x})]_T.$$

Example 2.4.8. Consider the differentiation operator $L : \mathbb{F}[x; 4] \rightarrow \mathbb{F}[x; 4]$ given by $L[p](x) = p'(x)$. Using the basis $S = [1, x, x^2, x^3, x^4]$ for both the domain and codomain, the matrix representation of L is

$$C_{SS} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

If we write $p(x) = 3x^4 - 8x^3 + 6x^2 - 12x + 4$ in terms of the basis S , we can use C_{SS} to compute $[L[p]]_S$. We have $[L[p]]_S = C_{SS}[p]_S$, or

$$[L[p]]_S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ -12 \\ 6 \\ -8 \\ 3 \end{bmatrix} = \begin{bmatrix} -12 \\ 12 \\ -24 \\ 12 \\ 0 \end{bmatrix}.$$

This agrees with the direct calculation of $p'(x) = 12x^3 - 24x^2 + 12x - 12$, written in terms of the basis S .

Remark 2.4.9. Let $V = \mathbb{F}^n$. The standard basis is the set

$$S = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n] = [(1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, 0, \dots, 1)].$$

Hence, an n -tuple $(a_1, \dots, a_n)$ can be written $\mathbf{x} = a_1\mathbf{e}_1 + \dots + a_n\mathbf{e}_n$, or in matrix form as

$$[\mathbf{x}]_S = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

We often write vectors in column form and suppress the notation $[\cdot]_S$ if the choice of basis is clear.

Remark 2.4.10. Unless the vector space and basis are already otherwise defined, when using matrices to represent linear operators, we assume that a matrix defines a linear operator on $\mathbb{F}^n$ with the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$.

2.5 Composition, Change of Basis, and Similarity

In the previous section, we showed how to represent a linear transformation as a matrix, once ordered bases for the domain and codomain are chosen. In this section, we first show that the composition of two linear transformations is represented by the matrix product, and then we use this to show how a matrix representation changes as a result of a change of basis.

2.5.1 Composition of Linear Transformations

We have seen that matrix-vector multiplication describes how linear transformations map vectors, and thus it should come as no surprise that composition of linear transformations corresponds to matrix-matrix multiplication. In this subsection we prove this result. A key consequence of this is that most questions in finite-dimensional linear algebra can be reformulated as problems in matrix analysis.

Theorem 2.5.1 (Matrix Multiplication). *Consider the vector spaces V , W , and X , with bases $S = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m]$, $T = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n]$, and $U = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p]$, respectively, together with the linear maps $L : V \rightarrow W$ and $K : W \rightarrow X$. Denote the*

unique matrix representations of L and K , respectively, as the $n \times m$ matrix $B_{TS} = [b_{jk}]$ and the $p \times n$ matrix $C_{UT} = [c_{ij}]$. Denote the unique matrix representation of the composition KL by $D_{US} = [d_{ij}]$. Matrix multiplication of the representation of K with the representation of L gives the representation of KL ; that is,

$$D_{US} = C_{UT}B_{TS}.$$

Proof. We begin by writing $\mathbf{v} \in V$ as $\mathbf{v} = \sum_{k=1}^m a_k \mathbf{s}_k$, together with $L(\mathbf{s}_k) = \sum_{j=1}^n b_{jk} \mathbf{t}_j$ and $K(\mathbf{t}_j) = \sum_{i=1}^p c_{ij} \mathbf{u}_i$. We have

$$\begin{aligned} KL(\mathbf{v}) &= K\left(\sum_{k=1}^m a_k L(\mathbf{s}_k)\right) = K\left(\sum_{k=1}^m a_k \sum_{j=1}^n b_{jk} \mathbf{t}_j\right) = \sum_{j=1}^n \left(\sum_{k=1}^m b_{jk} a_k\right) K(\mathbf{t}_j) \\ &= \sum_{j=1}^n \left(\sum_{k=1}^m b_{jk} a_k\right) \left(\sum_{i=1}^p c_{ij} \mathbf{u}_i\right) = \sum_{i=1}^p \left(\sum_{k=1}^m \left[\sum_{j=1}^n c_{ij} b_{jk}\right] a_k\right) \mathbf{u}_i. \end{aligned}$$

Thus, setting $d_{ik} = \sum_{j=1}^n c_{ij} b_{jk}$, we have

$$KL(\mathbf{v}) = \sum_{i=1}^p \left(\sum_{k=1}^m d_{ik} a_k\right) \mathbf{u}_i.$$

So $[KL(\mathbf{v})]_U = D_{US}[\mathbf{v}]_S$, and D_{US} is the matrix for the transformation KL . In matrix notation, this gives

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \cdots & d_{pm} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix},$$

which we may write as $D_{US} = C_{UT}B_{TS}$. $\square$

Remark 2.5.2. This result tells us, among other things, that the matrix representation of a linear transformation is an invertible matrix precisely when the linear transformation is invertible. In this case, we say that the matrix is *nonsingular*. If the matrix (or the corresponding transformation) is not invertible, we say it is *singular*.

Corollary 2.5.3. *Matrix multiplication is associative. That is, if $A \in M_{m \times n}$, $B \in M_{n \times k}$, and $C \in M_{k \times s}$, then we have*

$$(AB)C = A(BC).$$

Proof. First observe that function composition is associative; that is, given any functions $f : V \rightarrow W$, $g : W \rightarrow X$ and $h : X \rightarrow Y$, we have $(h \circ (g \circ f))(v) = h((g \circ f)(v)) = h(g(f(v))) = (h \circ g)(f(v)) = ((h \circ g) \circ f)(v)$ for every $v \in V$. The corollary follows since matrix multiplication is function composition. $\square$

2.5.2 Change of Basis and Transition Matrices

The results of the previous section make it possible for us to describe how changing the basis of the domain and codomain changes the matrix representation of a linear transformation.

Theorem 2.5.4. *Let C_{TS} be the matrix representation of $L : V \rightarrow W$ from the basis S into the basis T . If $P_{S\hat{S}}$ is the transition matrix on V from the basis $\hat{S}$ into S and $Q_{\hat{T}T}$ is the transition matrix on W from the basis T into $\hat{T}$, then the matrix representation $B_{\hat{T}\hat{S}}$ of L in terms of $\hat{S}$ and $\hat{T}$ is given by $B_{\hat{T}\hat{S}} = Q_{\hat{T}T}C_{TS}P_{S\hat{S}}$.*

Proof. Let $[y]_T = C_{TS}[x]_S$. We have $[x]_S = P_{S\hat{S}}[x]_{\hat{S}}$ and $[y]_{\hat{T}} = Q_{\hat{T}T}[y]_T$. When we combine these expressions we get $[y]_{\hat{T}} = Q_{\hat{T}T}C_{TS}P_{S\hat{S}}[x]_{\hat{S}}$. By uniqueness of matrix representations (Theorem 2.4.4), we know $B_{\hat{T}\hat{S}} = Q_{\hat{T}T}C_{TS}P_{S\hat{S}}$. $\square$

Remark 2.5.5. The following commutative diagram¹² illustrates the previous theorem:

$$\begin{array}{ccc}
 [x]_S & \xrightarrow{\text{left mult. by } C_{TS}} & [y]_T \\
 \uparrow \text{left mult. by } P_{S\hat{S}} & & \downarrow \text{left mult. by } Q_{\hat{T}T} \\
 [x]_{\hat{S}} & \xrightarrow{\text{left mult. by } B_{\hat{T}\hat{S}}} & [y]_{\hat{T}}
 \end{array}$$

Corollary 2.5.6. *Let S and $\hat{S}$ be bases for the vector space V with transition matrix $P_{S\hat{S}}$ from $\hat{S}$ into S . Let C_{SS} be the matrix representation of the operator $L : V \rightarrow V$ in the basis S . The matrix $B_{\hat{S}\hat{S}} = (P_{S\hat{S}})^{-1}C_{SS}P_{S\hat{S}}$ is the unique matrix representation of L in the basis $\hat{S}$.*

Remark 2.5.7. When the bases we are working with are understood from the context, we often drop the basis subscripts. In these cases, we usually also drop the square brackets $[\cdot]$ and denote a vector as an n -tuple $\mathbf{x} = (x_1, \dots, x_n)$ of its coordinates. However, when multiple bases are being considered in the same problem, it is wise to be very explicit with bases and write out all the basis subscripts so as to avoid confusion or mistakes.

Definition 2.5.8. *Two square matrices $A, B \in M_n(\mathbb{F})$ are similar if there exists a nonsingular $P \in M_n(\mathbb{F})$ such that $B = P^{-1}AP$.*

Remark 2.5.9. Any such P in the definition of similar matrices is a transition matrix defining a change of basis in $\mathbb{F}^n$, so two matrices are similar if and only if they correspond to the matrix representations, in different bases, of the same linear transformation $L : \mathbb{F}^n \rightarrow \mathbb{F}^n$.

¹²Starting at the bottom left corner of the commutative diagram and following the arrows up and around and back down to the bottom right gives the same result as following the single arrow across the bottom.

Remark 2.5.10. Similarity is an equivalence relation on the set of square matrices; see Exercise 2.27.

Remark 2.5.11. It is common to talk about the rank and nullity of a *matrix* when one really means the rank and nullity of the *linear transformation* represented by the matrix. Since any two similar matrices represent the same linear transformation (but with different bases), they must have the same rank and nullity.

Many other common properties of matrices are actually properties of the linear transformations that the matrices describe (for example, the determinant and the eigenvalues), both of which are discussed later. All of these properties must be the same for any two similar matrices.

Example 2.5.12. Let B be the unique matrix representation of the linear operator $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ in the basis $S = [\mathbf{s}_1, \mathbf{s}_2]$ given by

$$B = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix},$$

and let $\hat{S} = [\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2]$ be another basis for $\mathbb{R}^2$, defined by $\hat{\mathbf{s}}_1 = \mathbf{s}_1 - 2\mathbf{s}_2$ and $\hat{\mathbf{s}}_2 = \mathbf{s}_1$. The transition matrix P from $\hat{S}$ into S and its inverse are given by

$$P = \begin{bmatrix} 1 & 1 \\ -2 & 0 \end{bmatrix} \quad \text{and} \quad P^{-1} = \frac{1}{2} \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}.$$

Hence,

$$D = P^{-1}BP = \frac{1}{2} \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

which defines the linear transformation L in the basis $\hat{S}$. Thus, D is similar to B .

2.6 Important Example: Bernstein Polynomials

In this section we give an extended example of the concepts we have developed so far. Specifically, we define the Bernstein polynomials and show that they form bases for the spaces $\mathbb{F}[x; n]$ of polynomials. We then describe various linear transformations in terms of these bases and compare the results, via similarity, to the standard polynomial bases. These polynomials provide a useful tool for approximation and also play a key role in graphics applications, especially in the construction of Bézier curves.

In Volume 2 we also use the Bernstein polynomials to give a constructive proof of the Stone–Weierstrass approximation theorem, which guarantees that any continuous function on a closed interval can be approximated as closely as desired by polynomials.

Definition 2.6.1. Given $n \in \mathbb{N}$, the Bernstein polynomials $\{B_j^n(x)\}_{j=0}^n$ of degree n are defined as

$$B_j^n(x) = \binom{n}{j} x^j (1-x)^{n-j}, \quad \text{where} \quad \binom{n}{j} = \frac{n!}{j!(n-j)!}. \quad (2.6)$$

Remark 2.6.2. Observe that $B_0^n(0) = 1$, and $B_j^n(0) = 0$ if $j \neq 0$. Moreover, $B_n^n(1) = 1$, and $B_j^n(1) = 0$ if $j \neq n$.

The first interesting property of the Bernstein polynomials is that they sum to 1, as can be seen from the binomial theorem:

$$\sum_{j=0}^n B_j^n = \sum_{j=0}^n \binom{n}{j} x^j (1-x)^{n-j} = (x + (1-x))^n = 1.$$

We show that the set $\{B_j^n(x)\}_{j=0}^n$ forms a basis for $\mathbb{F}[x; n]$ and provides the transition matrix between the Bernstein basis and the usual monomial basis $[1, x, x^2, \dots, x^n]$.

Example 2.6.3. The Bernstein polynomials of degree $n = 4$ are

$$\begin{aligned} B_0^4(x) &= (1-x)^4, & B_1^4(x) &= 4x(1-x)^3, & B_2^4(x) &= 6x^2(1-x)^2, \\ B_3^4(x) &= 4x^3(1-x), & B_4^4(x) &= x^4. \end{aligned}$$

These are plotted in Figure 2.2.

We compute the transition matrix P_{ST} from the Bernstein basis

$$T = [B_0^4(x), B_1^4(x), B_2^4(x), B_3^4(x), B_4^4(x)]$$

into the standard polynomial basis $S = [1, x, x^2, x^3, x^4]$, and its inverse $Q_{TS} = (P_{ST})^{-1}$, as

$$P_{ST} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -4 & 4 & 0 & 0 & 0 \\ 6 & -12 & 6 & 0 & 0 \\ -4 & 12 & -12 & 4 & 0 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix} \quad \text{and} \quad Q_{TS} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1/4 & 0 & 0 & 0 \\ 1 & 1/2 & 1/6 & 0 & 0 \\ 1 & 3/4 & 1/2 & 1/4 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Let $p(x) = 3x^4 - 8x^3 + 6x^2 - 12x + 4$. By computing $[p(x)]_T = Q_{TS}[p(x)]_S$, we see that $p(x)$ can be written in the Bernstein basis as

$$p(x) = 4B_0^4(x) + B_1^4(x) - B_2^4(x) - 4B_3^4(x) - 7B_4^4(x).$$

Lemma 2.6.4. We have the following identity for $j = 0, 1, \dots, n$:

$$B_j^n(x) = \sum_{i=j}^n (-1)^{i-j} \binom{n}{i} \binom{i}{j} x^i. \quad (2.7)$$

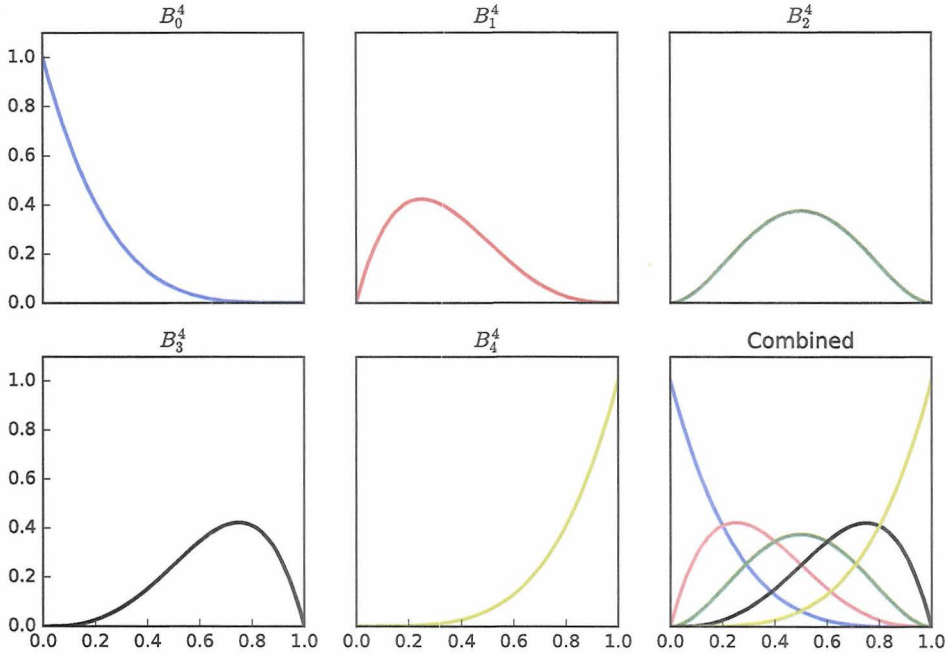


Figure 2.2. The five Bernstein polynomials of degree 4 on the interval $[0, 1]$, plotted first separately and then together (bottom right). See also Example 2.6.3.

Proof.

$$\begin{aligned}
 B_j^n(x) &= \binom{n}{j} x^j (1-x)^{n-j} = \binom{n}{j} x^j \sum_{k=0}^{n-j} \binom{n-j}{k} (-1)^k x^k \\
 &= \sum_{k=0}^{n-j} (-1)^k \binom{n}{j} \binom{n-j}{k} x^{j+k} = \sum_{i=j}^n (-1)^{i-j} \binom{n}{j} \binom{n-j}{i-j} x^i \\
 &= \sum_{i=j}^n (-1)^{i-j} \frac{n!}{j!(n-j)!} \frac{(n-j)!}{(i-j)!(n-j-i+j)!} x^i \\
 &= \sum_{i=j}^n (-1)^{i-j} \frac{n! i!}{j! (i-j)! (n-i)! i!} x^i \\
 &= \sum_{i=j}^n (-1)^{i-j} \binom{n}{i} \binom{i}{j} x^i. \quad \square
 \end{aligned}$$

Theorem 2.6.5. For any $n \in \mathbb{N}$ the set $T_n = \{B_j^n(x)\}_{j=0}^n$ of degree- n Bernstein polynomials forms a basis for $\mathbb{F}[x; n]$.

Proof. To see that T_n is a basis, we construct (what will be transition) matrices P and Q such that $Q = P^{-1}$ and such that any linear relation $\sum_{i=0}^n c_i B_i^n(x) = 0$ can be written as

$$[1, x, \dots, x^n] P [c_0, \dots, c_n]^T = 0.$$

Since $S = [1, x, \dots, x^n]$ is a basis for $\mathbb{F}[x; n]$, this means that the vector $P[c_0, \dots, c_n]^T$ must vanish; that is, $P[c_0, \dots, c_n]^T = \mathbf{0}$. Since P is invertible, we have

$$[c_0, \dots, c_n]^T = (QP)[c_0, \dots, c_n]^T = Q(P[c_0, \dots, c_n]^T) = Q\mathbf{0} = \mathbf{0}.$$

Therefore the set of $n+1$ polynomials $\{B_0^n(x), \dots, B_n^n(x)\}$ are linearly independent in the $(n+1)$ -dimensional space $\mathbb{F}[x; n]$, and hence form a basis for $\mathbb{F}[x; n]$.

We now construct the matrices P and Q . Let $P = [p_{jk}]$ be the $(n+1) \times (n+1)$ lower-triangular matrix defined by

$$p_{jk} = \begin{cases} (-1)^{j-k} \binom{n}{j} \binom{j}{k} & \text{if } j \geq k \\ 0 & \text{if } j < k \end{cases} \quad \text{for } j, k \in \{0, 1, \dots, n\}. \quad (2.8)$$

If we already knew that the Bernstein polynomials formed a basis, then (2.7) tells us that P would be the transition matrix P_{ST} from the Bernstein basis T_n into the power basis S .

Now define the matrix $Q = [q_{ij}]$ by

$$q_{ij} = \begin{cases} \binom{i}{j} / \binom{n}{j} & \text{if } i \geq j \\ 0 & \text{if } i < j \end{cases} \quad \text{for } i, j \in \{0, 1, \dots, n\}. \quad (2.9)$$

Once we see that the set T_n of Bernstein polynomials forms a basis, it will be clear that matrix Q is the transition matrix Q_{TS} from the basis S to the basis T_n .

We verify that $QP = I = PQ$ by direct computation. When $0 \leq k \leq i \leq n$, the product QP takes the form

$$(QP)_{ik} = \sum_{j=0}^n q_{ij} p_{jk} = \sum_{j=k}^i q_{ij} p_{jk} = \sum_{j=k}^i (-1)^{j-k} \binom{i}{j} \binom{j}{k} = B_k^i(1).$$

When $i = k$ we have that $B_k^i(1) = 1$, and when $k < i$ we have that $B_k^i(1) = 0$. Also, when $0 \leq i < k \leq n$, we have that $(QP)_{ik} = 0$ since both matrices are lower-triangular. Hence, $QP = I$, and by Proposition 2.3.17 we also have $PQ = I$.

Thus, P and Q have the necessary properties, and the argument outlined above shows that $\{B_j^n(x)\}_{j=0}^n$ is a basis of $\mathbb{F}[x; n]$. $\square$

Example 2.6.6. Using the matrix A in Example 2.4.8, representing the derivative of a polynomial in $\mathbb{F}[x; 4]$, we use similarity to transform it to the Bernstein basis. Multiplying out $P_{ST}^{-1} A P_{ST}$, we get

$$P_{ST}^{-1}AP_{ST} = \begin{bmatrix} -4 & 4 & 0 & 0 & 0 \\ -1 & -2 & 3 & 0 & 0 \\ 0 & -2 & 0 & 2 & 0 \\ 0 & 0 & -3 & 2 & 1 \\ 0 & 0 & 0 & -4 & 4 \end{bmatrix}.$$

Thus, using the representation of $p(x)$ in Example 2.6.3, we express the derivative as

$$\begin{bmatrix} -4 & 4 & 0 & 0 & 0 \\ -1 & -2 & 3 & 0 & 0 \\ 0 & -2 & 0 & 2 & 0 \\ 0 & 0 & -3 & 2 & 1 \\ 0 & 0 & 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ -1 \\ -4 \\ -7 \end{bmatrix} = \begin{bmatrix} -12 \\ -9 \\ -10 \\ -12 \\ -12 \end{bmatrix}.$$

Thus, $p'(x) = -12B_0^n(x) - 9B_1^n(x) - 10B_2^n(x) - 12B_3^n(x) - 12B_4^n(x)$.

Application 2.6.7 (Computer Aided Design). Over a century ago, when Bernstein first introduced his polynomials, they were seen as being mostly a theoretical construct with limited application. However, in the early 1960s two French engineers, de Casteljau and Bézier, independently used Bernstein polynomials to address a perplexing problem in the automotive design industry, namely, how to specify “free-form” shapes consistently. Their solution was to use Bernstein polynomials to describe parametric curves and surfaces with a finite number of points that control the shape of the curve or surface. This enabled the use of computers to sketch the curves and surfaces. The resulting curves are called *Bézier curves* and are of the form

$$r(t) = \sum_{j=0}^n \mathbf{p}_j B_j^n(t), \quad t \in [0, 1],$$

where $\mathbf{p}_0, \dots, \mathbf{p}_n$ are the control points and $B_j^n(t)$ are the Bernstein polynomials.

2.7 Linear Systems

In this section we discuss the theory behind solving linear systems. We develop the theoretical foundation of row reduction, and then we show how to find bases for the kernel and range of a linear transformation.

Consider the linear system $A\mathbf{x} = \mathbf{b}$, where $A \in M_{m \times n}(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^m$ are given and $\mathbf{x} \in \mathbb{F}^n$ is unknown. We have already seen that a solution exists if and only if $\mathbf{b} \in \mathcal{R}(A)$. Moreover, if a solution exists and $\mathcal{N}(A) \neq \{\mathbf{0}\}$, then there are infinitely many solutions.

Knowing that a solution exists is usually not sufficient; we often want to know what that solution is. The standard way to solve a linear system by hand is *row reduction*. Row reduction is the process of performing *row operations* to transform

the linear system into a form that is easy to solve, either directly by inspection or through a process called *back substitution*. In this section, we explore row reduction in greater depth.

2.7.1 Elementary Matrices

A row operation is performed by left multiplying both sides of a linear system by an *elementary matrix*, as described below. There are three types of row operations.

Type I: Swapping rows, denoted $R_i \leftrightarrow R_j$. The corresponding elementary matrix (called a type I matrix) is formed by interchanging rows i and j of the identity matrix. Left multiplying by this matrix performs the row operation. For example,

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} d & e & f \\ a & b & c \\ g & h & i \end{bmatrix}.$$

A type I matrix is also called a transposition matrix.

Type II: Multiply a row by a nonzero scalar α , denoted $R_j \rightarrow \alpha R_j$. A type II elementary matrix is formed by multiplying the j th diagonal entry of the identity matrix by α . For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & b & c \\ \alpha d & \alpha e & \alpha f \\ g & h & i \end{bmatrix}.$$

Type III: Add a scalar multiple of a row to another row, denoted $R_i \rightarrow \alpha R_j + R_i$ (with $i \neq j$). A type III elementary matrix is formed by inserting α in the (i, j) entry of the identity matrix. For example, if $(i, j) = (2, 1)$, then we have

$$\begin{bmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & b & c \\ \alpha a + d & \alpha b + e & \alpha c + f \\ g & h & i \end{bmatrix}.$$

Remark 2.7.1. All three elementary matrices are found by performing the desired row operation on the identity matrix.

Proposition 2.7.2. *All of the elementary matrices are invertible.*

Proof. This is easily verified by direct computation: If E is type I, then $E^2 = I$. If E is type II corresponding to $R_j \rightarrow \alpha R_j$, then its inverse corresponds to $R_j \rightarrow \alpha^{-1} R_j$ (recall $\alpha \neq 0$). Finally, if E is type III, corresponding to $R_i \rightarrow \alpha R_j + R_i$, then its inverse corresponds to $R_i \rightarrow -\alpha R_j + R_i$. $\square$

Remark 2.7.3. Since the elementary matrices are invertible, row reduction can be viewed as repeated left multiplication of both sides of a linear system by invertible matrices. Multiplication by invertible matrices represents a change of basis, and left multiplication by the elementary matrices can be thought of as simply changing

the basis of the range, but leaving the domain unchanged. Thus, any solution of the system $E\mathbf{A}\mathbf{x} = \mathbf{0}$ is also a solution of the system $\mathbf{A}\mathbf{x} = \mathbf{0}$, and conversely.

More generally, any solution of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is a solution of the system $E\mathbf{A}\mathbf{x} = E\mathbf{b}$ for any elementary matrix E , and conversely. This is equivalent to saying that a solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be found by row reducing both the matrix A and the vector $\mathbf{b}$ using the same operations in the same order. The goal is to judiciously choose these elementary matrices so that the linear system is reduced to some nice form that is easy to solve. Typically this is either an upper-triangular matrix or a diagonal matrix.

Example 2.7.4. Consider the linear system

$$\begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}.$$

To solve it using row operations, or equivalently elementary matrices, we first swap the rows. This is done by left multiplying both sides by the type I matrix corresponding to $R_1 \leftrightarrow R_2$. Thus, we have

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}.$$

Now we eliminate the $(2, 1)$ entry by using the type III matrix corresponding to $R_2 \rightarrow -2R_1 + R_2$. This yields

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ -5 \end{bmatrix}. \quad (2.10)$$

From here we see that $x_2 = -5$ and $x_1 + x_2 = 4$, which reduces to $x_1 = 9$ by back substitution. Alternatively, we can continue by eliminating the $(1, 2)$ entry using the type III matrix $R_1 \rightarrow -R_2 + R_1$. Thus, we have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -5 \end{bmatrix} = \begin{bmatrix} 9 \\ -5 \end{bmatrix}.$$

Remark 2.7.5. As a shorthand in the previous example, we write the original linear system in *augmented* form

$$\left[\begin{array}{cc|c} 2 & 3 & 3 \\ 1 & 1 & 4 \end{array} \right].$$

Thus, we can more compactly carry out the row reduction process by writing

$$\left[\begin{array}{cc|c} 2 & 3 & 3 \\ 1 & 1 & 4 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 1 & 4 \\ 2 & 3 & 3 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 1 & 4 \\ 0 & 1 & -5 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 0 & 9 \\ 0 & 1 & -5 \end{array} \right].$$

Definition 2.7.6. The matrix B is row equivalent to the matrix A if there exists a finite collection of elementary matrices $E_1, E_2, \dots, E_n$ such that $B = E_1 E_2 \cdots E_n A$.

Theorem 2.7.7. Row equivalence is an equivalence relation.

Proof. See Exercise 2.40. $\square$

Remark 2.7.8. Sometimes we say that one matrix is row equivalent to another matrix, but more commonly we say that one augmented matrix is row equivalent to another augmented matrix, or, in other words, that one linear system is row equivalent to another linear system. Thus, if two linear systems (or augmented matrices) are row equivalent, they have the same solution.

2.7.2 Row Echelon Form

Definition 2.7.9. A matrix A is in row echelon form (REF) if the following hold:

- (i) The first nonzero entry, called the leading entry, of each nonzero row is always strictly to the right of the leading entry of the row above it.
- (ii) All nonzero rows are above any zero rows.

A matrix A is in reduced row echelon form (RREF) if the following hold:

- (i) It is in REF.
- (ii) The leading entry of every nonzero row is equal to one.
- (iii) The leading entry of every nonzero row is the only nonzero entry in its column.

Example 2.7.10. The first two of the following matrices are in REF, the next is in RREF, and the last two are not in REF. Can you say why?

$$\begin{bmatrix} 2 & 2 & 3 & 2 \\ 0 & 4 & 5 & 3 \\ 0 & 0 & 6 & 7 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 4 & 2 & 2 & 1 \\ 0 & 3 & 4 & 5 & 0 \\ 0 & 0 & 0 & 6 & 9 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 4 & 0 & 2 \\ 0 & 1 & 7 & 0 & 9 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 4 \\ 0 & 5 & 6 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 0 \\ 0 & 0 & 6 \end{bmatrix}.$$

Remark 2.7.11. REF is the canonical form of a linear system that can then be solved by what we call *back substitution*. In other words, when an augmented matrix is in REF, we can solve for the coordinates of $\mathbf{x}$ and substitute them back into the partial solution, as we did with (2.10).

Nota Bene 2.7.12. Although it is much easier to determine the solution of a linear system when it is in RREF than REF, it usually requires twice as many row operations to get a system into RREF, and so it is faster algorithmically to reduce to REF and then use back substitution.

Proposition 2.7.13. *If A is a matrix, then there exists an RREF matrix B such that A and B are row equivalent. Moreover, if A is a square matrix, then B is upper triangular (the (i, j) entry of B is zero whenever $i > j$). Finally, if T is a square, upper-triangular matrix, all of whose diagonal elements are nonzero, then T is row equivalent to the identity matrix.*

Proof. See Exercise 2.39. $\square$

Theorem 2.7.14. *A square matrix is nonsingular if and only if it is row equivalent to the identity matrix.*

Proof. ($\implies$) By the previous proposition, any square matrix A can be row reduced to an RREF upper-triangular matrix B by a sequence $E_1, E_2, \dots, E_k$ of elementary matrices such that $E_1 E_2 \cdots E_k A = B$. If A is nonsingular, then it has an inverse A^{-1} , and so

$$I = E_1 E_2 \cdots E_k A A^{-1} (E_1 E_2 \cdots E_k)^{-1} = B A^{-1} (E_1 E_2 \cdots E_k)^{-1}.$$

Assume by way of contradiction that one of the diagonal elements of B is zero. Since B is in RREF, the entire bottom row of B must be zero. This implies that the bottom row of the product $B A^{-1} (E_1 E_2 \cdots E_k)^{-1}$ is zero, and hence it is not equal to I —a contradiction.

Therefore B can have no zeros on the diagonal and must be row equivalent to I by Proposition 2.7.13. Since B is row equivalent to I , then A is also row equivalent to I .

($\impliedby$) If A is row equivalent to the identity, then there exists a sequence $E_1, E_2, \dots, E_k$ of elementary matrices such that $E_1 E_2 \cdots E_k A = I$. It follows that $A^{-1} = E_1 E_2 \cdots E_k$. $\square$

Application 2.7.15 (LU Decomposition). There are efficient numerical libraries for solving linear systems of the form $A\mathbf{x} = \mathbf{b}$. If nothing is known about the matrix A , or if A has no specialized structure, the best known routine is called *LU factorization*, which produces two output matrices, specifically, an upper-triangular matrix U which corresponds to A in REF form, and a lower-triangular matrix L which is the product of type III elementary matrices. Specifically, if $E_n \cdots E_1 A = U$, then $L = (E_n \cdots E_1)^{-1} = E_1^{-1} E_2^{-1} \cdots E_n^{-1}$. The inverse of a type III elementary matrix is a type III elementary matrix, and so L is indeed the product of type III matrices.

To solve a linear system $A\mathbf{x} = \mathbf{b}$, first write A as $A = LU$, and then solve for $\mathbf{y}$ in the system $L(\mathbf{y}) = \mathbf{b}$ via back substitution. Finally, solve for $\mathbf{x}$ in the system $U\mathbf{x} = \mathbf{y}$, again via back substitution.

Corollary 2.7.16. *If a matrix A is invertible, then the RREF of the augmented matrix $[A \mid I]$ is $[I \mid A^{-1}]$. In other words, we can compute A^{-1} by row reducing $[A \mid I]$.*

Proof. When we put the augmented matrix $[A \mid I]$ in RREF using a sequence of elementary matrices $E_1, E_2, \dots, E_k$, we have $E_1 E_2 \cdots E_k [A \mid I] = [I \mid E_1 E_2 \cdots E_k]$. This shows that $E_1 E_2 \cdots E_k A = I$ and hence $E_1 E_2 \cdots E_k = A^{-1}$. $\square$

2.7.3 Basic and Free Variables

In the absence of special structure, row reduction is the canonical approach¹³ to computing the solution of the linear system $A\mathbf{x} = \mathbf{b}$, where $A \in M_{m \times n}(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^m$. If A is invertible, then row reduction is straightforward. If A is singular, and in particular if it is rectangular, then the problem is a little more complicated.

Consider first the homogeneous linear system $A\mathbf{x} = \mathbf{0}$. For any finite sequence of $m \times m$ elementary matrices $E_1, \dots, E_k$, the system has the same solution as the row-reduced system $E_k \cdots E_1 A\mathbf{x} = E_k \cdots E_1 \mathbf{0} = \mathbf{0}$. Thus, row reduction does not change the solution $\mathbf{x}$, which means that the kernel of A is the same as the kernel of the RREF matrix $B = E_k \cdots E_1 A$.

Given an RREF matrix B , we associate the columns with leading entries as the *basic variables* and the ones without as *free variables*. The number of free variables is the dimension of the kernel and corresponds to the number of degrees of freedom of the kernel. More precisely, the basic variables can be written as linear combinations of the free variables, and thus the entire solution set to the homogeneous problem can be written as a linear combination of the free variables (one for each dimension).

If the dimension of the kernel equals the number of free variables, then by the rank-nullity theorem (Corollary 2.3.9), the number of basic variables is equal to the rank of the matrix.

Example 2.7.17. Consider the linear system $A\mathbf{x} = \mathbf{0}$, where

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \quad (2.11)$$

¹³The widely used computing libraries for solving linear systems have many efficiencies and enhancements built in that are quite sophisticated. We present only the entry-level approach here.

We can write this as an augmented linear system and row reduce to RREF form (using elementary matrices)

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 5 & 6 & 7 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 2 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & 2 & 0 \end{array} \right].$$

Thus, the system is row equivalent to the reduced system $B\mathbf{x} = \mathbf{0}$, where

$$B = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \quad (2.12)$$

The first two columns correspond to basic variables (x_1 and x_2), and the third column corresponds to the free variable x_3 . The augmented system can then be written as

$$\begin{aligned} x_1 &= x_3, \\ x_2 &= -2x_3, \end{aligned}$$

or equivalently the solution is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_3 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

where $x_3 \in \mathbb{F}$. It follows that $\mathcal{N}(A) = \text{span}\{[1 \ -2 \ 1]^\top\}$. Here it is clear that the free column $[-1 \ 2]^\top$ can be written as a linear combination of the basic columns $[1 \ 0]^\top$ and $[0 \ 1]^\top$. Hence, the rank is 2 and the nullity is 1.

In the more general nonhomogeneous case of $A\mathbf{x} = \mathbf{b}$, the set of all solutions is a coset of the form $\mathbf{x}' + \mathcal{N}(A)$, where $\mathbf{x}'$ is any particular solution of $A\mathbf{x} = \mathbf{b}$. Since all the elementary matrices are invertible, solving $A\mathbf{x} = \mathbf{b}$ is equivalent to solving $B\mathbf{x} = \mathbf{d}$, where B is given above and $\mathbf{d} = E_k \cdots E_1 \mathbf{b}$. This is generally done by row reducing the augmented matrix $[A \mid \mathbf{b}]$. This allows us to track the effect of multiplying elementary matrices on both A and $\mathbf{b}$ simultaneously.

Example 2.7.18. Consider the linear system $A\mathbf{x} = \mathbf{b}$, where A is given in the previous example and $\mathbf{b} = [4 \ 8]^\top$. We can write this as an augmented linear system and row reduce to RREF form (using elementary matrices)

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \end{array} \right].$$

Thus, the system is row equivalent to the reduced system $B\mathbf{x} = \mathbf{d}$, where B is given in (2.12) and $\mathbf{d} = [-2 \ 3]^\top$. The augmented system can be written as

$$\begin{aligned}x_1 &= x_3 - 2, \\x_2 &= -2x_3 + 3, \\x_3 &= x_3,\end{aligned}$$

or equivalently as $\mathbf{x} = \mathbf{x}' + \mathcal{N}(A)$, where $\mathbf{x}' = [-2 \ 3 \ 0]^\top$ and $\mathcal{N}(A)$ is given in the previous example.

If we think of the matrices E_i as describing a change of basis, then they are changing the basis of the codomain, but they leave the domain unchanged. The columns of A are vectors in the codomain that span the range of A . The columns of B are those same vectors expressed in a new basis, and their span is the range of the same linear transformation, but expressed in this new basis. However, if B is in RREF, then it is easy to see that the columns with leading entries (called *basic columns*) span the range (again, expressed in terms of the new basis). This means that the corresponding columns of the original matrix A span the range (in terms of the original basis).

Example 2.7.19. In the previous two examples, the first and second columns are basic and obviously linearly independent, and the corresponding columns from the original matrix are also linearly independent and form a basis of the range, so we have

$$\mathcal{R}(A) = \text{span} \left\{ \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 6 \end{bmatrix} \right\}.$$

In summary, to find a basis for the kernel of a matrix A , row reduce A to RREF and write the basic variables in terms of the free variables. To find a general solution to $A\mathbf{x} = \mathbf{b}$, row reduce the augmented matrix $[A \mid \mathbf{b}]$ to RREF and use it to find a single solution $\mathbf{y}$ of the system. The coset $\mathbf{y} + \mathcal{N}(A)$ is the set of all solutions of $A\mathbf{x} = \mathbf{b}$. Finally, the range of A is the span of the columns, but these columns may not be linearly independent. To get a basis for the range of A , row reduce to RREF, and then the columns of A that correspond to basic variables form the basis for $\mathcal{R}(A)$.

2.8 Determinants I

In this section and the next, we develop the classical theory of determinants. Most linear algebra texts define the determinant using the cofactor expansion. This is a fair approach, but it is difficult to prove some of the properties of the determinant rigorously this way, and so a little hand waving is common. Instead, we develop the determinant using permutations. While this approach is initially a little more difficult, it allows us to give a rigorous derivation of the main properties of the determinant. Moreover, permutations are a powerful mathematical tool with broad applicability beyond determinants, and so we favor this approach pedagogically.

We begin by developing properties of permutations. We then prove two key properties of the determinant, specifically that the determinant of an upper-triangular matrix is the product of its diagonal elements and that the determinant of a matrix is equal to the determinant of its transpose.

Neither the cofactor definition nor the permutation definition are well suited to numerical computation. Indeed, the number of arithmetic operations required to compute the determinant of an $n \times n$ matrix, using either definition, is proportional to n factorial. In the following section, we introduce a much more efficient way of computing the determinant using row reduction. This provides an algorithm that grows in proportion to n^3 instead of $n!$.

Finally, we remark that while determinants are useful in developing theory, they are seldom used in computation because there are generally faster methods available. As a result there are schools of thought that eschew determinants altogether. One place where determinants are essential, however, is in describing how a linear transformation changes volumes. For example, given a parallelepiped spanned by n vectors in $\mathbb{R}^n$, the absolute value of the determinant of the matrix whose columns are these vectors is the volume of the parallelepiped. And a linear transformation L takes the unit n -cube spanned by the standard basis vectors to a parallelepiped spanned by the columns of the matrix representation of L . So, the transformation L changes the cube to a parallelepiped of volume $\det(L)$. This fact is essential to changing variables in multidimensional integration (see Section 8.7).

2.8.1 Permutations

Throughout this section, assume that n is a positive integer and $T_n = \{1, 2, \dots, n\}$.

Definition 2.8.1. A permutation σ of a set T_n is a bijection $\sigma : T_n \rightarrow T_n$. The set of all permutations of T_n is denoted S_n and is called the symmetric group of order n .

Remark 2.8.2. The composition of two permutations on a given set is again a permutation, and the inverse of every permutation is again a permutation. A nonempty set of functions is called a *group* if it has these two properties.

Notation 2.8.3. If $\sigma \in S_n$ is given by $\sigma(j) = i_j$ for every $j \in T_n$, then we denote σ by the ordered tuple $[[i_1, \dots, i_n]]$. For example, if $\sigma \in S_3$ is such that $\sigma(1) = 2$, $\sigma(2) = 3$, and $\sigma(3) = 1$, then we write σ as the 3-tuple $[[2, 3, 1]]$.

Example 2.8.4. We have

$$\begin{aligned} S_1 &= \{[[1]]\}, \\ S_2 &= \{[[1, 2]], [[2, 1]]\}, \\ S_3 &= \{[[1, 2, 3]], [[2, 3, 1]], [[3, 1, 2]], [[2, 1, 3]], [[1, 3, 2]], [[3, 2, 1]]\}. \end{aligned}$$

Note that $|S_1| = 1$, $|S_2| = 2$, and $|S_3| = 6$. For general values of n , we have the following theorem.

Theorem 2.8.5. *The number of permutations of the set T_n is $n!$; that is, $|S_n| = n!$.*

Proof. We proceed by induction on $n \in \mathbb{N}$. By the previous example, we know that the base case holds; thus we assume that $|S_{n-1}| = (n-1)!$. For each permutation $[\sigma(1), \sigma(2), \dots, \sigma(n-1)]$ in S_{n-1} , we can create n permutations of T_n as follows:

$$[n, \sigma(1), \dots, \sigma(n-1)], [\sigma(1), n, \dots, \sigma(n-1)], \dots, [\sigma(1), \dots, \sigma(n-1), n].$$

All permutations of T_n can be produced in this way, and no two are the same. Thus, there are $n \cdot (n-1)! = n!$ permutations in S_n . $\square$

Definition 2.8.6. *Let $\sigma : T_n \rightarrow T_n$ be a permutation. An inversion in σ is a pair $(\sigma(i), \sigma(j))$ such that $i < j$ and $\sigma(i) > \sigma(j)$.*

Definition 2.8.7. *A permutation σ is called even if it has an even number of inversions. It is called odd if it has an odd number of inversions. The sign of the permutation is defined as*

$$\text{sign } \sigma = \begin{cases} 1 & \text{if } \sigma \text{ is even,} \\ -1 & \text{if } \sigma \text{ is odd.} \end{cases}$$

Example 2.8.8. The permutation $\sigma = [3, 2, 4, 5, 1]$ on T_5 has 5 inversions, namely, $(3, 2)$, $(3, 1)$, $(2, 1)$, $(4, 1)$, and $(5, 1)$. Thus, σ is an odd permutation.

Example 2.8.9. A permutation that swaps two elements, fixing the rest, is a *transposition*. More precisely, σ is a transposition of T_n if there exists $i, j \in T_n$, $i \neq j$, such that $\sigma(i) = j$, $\sigma(j) = i$, and $\sigma(k) = k$ for every $k \neq i, j$. For example, the permutations $[2, 1, 3, 4, 5]$ and $[1, 2, 5, 4, 3]$ are both transpositions of T_5 .

Assuming $i < j$, the inversions of the transposition are (j, i) , $(j, i+1)$, $(j, i+2), \dots, (j, j-1)$ and $(i+1, i)$, $(i+2, i), \dots, (j-1, i)$. This adds up to $2(j-i)-1$, which is odd. Hence, the sign of the transposition is -1 .

Lemma 2.8.10 (Equivalent Definition of $\text{sign}(\sigma)$). *Let $\sigma \in S_n$. Consider the polynomial $p(x_1, x_2, \dots, x_n) = \prod_{i < j} (x_i - x_j)$. The sign of a permutation is given as*

$$\text{sign}(\sigma) = \frac{p(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)})}{p(x_1, x_2, \dots, x_n)}. \quad (2.13)$$

Proof. It is straightforward to see that (2.13) reduces to $(-1)^m$, where m is the number of inversions, and thus even (respectively, odd) permutations are positive (respectively, negative). $\square$

Example 2.8.11. When $n = 4$ and $\sigma = [2, 4, 3, 1]$, we have

$$\begin{aligned} \text{sign}(\sigma) &= \frac{p(x_{\sigma(1)}, x_{\sigma(2)}, x_{\sigma(3)}, x_{\sigma(4)})}{p(x_1, x_2, x_3, x_4)} \\ &= \frac{p(x_2, x_4, x_3, x_1)}{p(x_1, x_2, x_3, x_4)} \\ &= \frac{(x_2 - x_4)(x_2 - x_3)(x_2 - x_1)(x_4 - x_3)(x_4 - x_1)(x_3 - x_1)}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)} \\ &= 1 \cdot 1 \cdot -1 \cdot -1 \cdot -1 \cdot -1 = 1 = (-1)^4. \end{aligned}$$

Observe that there are four inversions, so $\text{sign}(\sigma) = 1$.

The next theorem is an important tool for working with signs.

Theorem 2.8.12. *The sign of the composition of two permutations is equal to the product of their signs; that is, if $\sigma, \tau \in S_n$, then $\text{sign}(\sigma \circ \tau) = \text{sign}(\sigma) \text{sign}(\tau)$.*

Proof. Note that $x_1, x_2, \dots, x_n$ in (2.13) are dummy variables, so we could just as well have used $y_n, y_{n-1}, \dots, y_1$ or any other set of variables in any other initial order. Thus, for any $\sigma, \tau \in S_n$ we have

$$\text{sign}(\sigma) = \frac{p(x_{\sigma(\tau(1))}, x_{\sigma(\tau(2))}, \dots, x_{\sigma(\tau(n))})}{p(x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)})}.$$

Computing the sign of $\sigma\tau$ using (2.13), we have

$$\begin{aligned} \text{sign}(\sigma\tau) &= \frac{p(x_{\sigma(\tau(1))}, x_{\sigma(\tau(2))}, \dots, x_{\sigma(\tau(n))})}{p(x_1, x_2, \dots, x_n)} \\ &= \frac{p(x_{\sigma(\tau(1))}, x_{\sigma(\tau(2))}, \dots, x_{\sigma(\tau(n))})}{p(x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)})} \cdot \frac{p(x_{\tau(1)}, x_{\tau(2)}, \dots, x_{\tau(n)})}{p(x_1, x_2, \dots, x_n)} \\ &= \text{sign}(\sigma) \text{sign}(\tau). \quad \square \end{aligned}$$

Corollary 2.8.13. *If σ is a permutation, then $\text{sign}(\sigma^{-1}) = \text{sign}(\sigma)$.*

Proof. If e is the identity map, then $1 = \text{sign}(e) = \text{sign}(\sigma\sigma^{-1}) = \text{sign}(\sigma) \text{sign}(\sigma^{-1})$. Thus, if σ is even (respectively, odd), then so is σ^{-1} . $\square$

2.8.2 Definition of the Determinant

Definition 2.8.14. *Let $A = [a_{ij}]$ be in $M_n(\mathbb{F})$ and $\sigma \in S_n$. The elementary product associated with $\sigma \in S_n$ is the product $a_{1\sigma(1)}a_{2\sigma(2)} \cdots a_{n\sigma(n)}$.*

Remark 2.8.15. An elementary product contains exactly one element from each row of A and exactly one element of each column.

Example 2.8.16. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad (2.14)$$

and let $\sigma \in S_3$ be given by $[[3, 1, 2]]$. The elementary product is $a_{13}a_{21}a_{32}$. Table 2.1 contains a complete list of permutations and their corresponding elementary products for a 3×3 matrix.

Definition 2.8.17. If $A = [a_{ij}] \in M_n(\mathbb{F})$, then the determinant of A is the quantity

$$\det(A) = \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}.$$

Example 2.8.18. Given $n = 2$, there are two permutations on T_2 , namely, $\sigma_1 = [[1, 2]]$ and $\sigma_2 = [[2, 1]]$. Since $\text{sign}(\sigma_1) = 1$ and $\text{sign}(\sigma_2) = -1$, it follows that the determinant of the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

is given by $\det(A) = a_{11}a_{22} - a_{12}a_{21}$.

σ	$\text{sign}(\sigma)$	elementary product
$[[1, 2, 3]]$	1	$a_{11}a_{22}a_{33}$
$[[1, 3, 2]]$	-1	$a_{11}a_{23}a_{32}$
$[[2, 3, 1]]$	1	$a_{12}a_{23}a_{31}$
$[[2, 1, 3]]$	-1	$a_{12}a_{21}a_{33}$
$[[3, 1, 2]]$	1	$a_{13}a_{21}a_{32}$
$[[3, 2, 1]]$	-1	$a_{13}a_{22}a_{31}$

Table 2.1. A list of the permutations, their signs, and the corresponding elementary products for a generic 3×3 matrix.

Example 2.8.19. Let A be the arbitrary 3×3 matrix as given in (2.14). Using Table 2.1, we have

$$\det(A) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

2.8.3 Elementary Properties of Determinants

Theorem 2.8.20. *The determinant of an upper-triangular matrix is the product of its diagonals; that is, $\det(A) = \prod_{i=1}^n a_{ii}$.*

Proof. Since $A = [a_{ij}]$ is upper triangular, then $a_{ij} = 0$ whenever $i > j$. Thus, for an elementary product $a_{1\sigma(1)}a_{2\sigma(2)} \cdots a_{n\sigma(n)}$ to be nonzero, it is necessary that $\sigma(k) \geq k$ for each k . But $\sigma(n) \geq n$ implies that $\sigma(n) = n$. Similarly, $\sigma(n-1) \geq n-1$, but since $\sigma(n) = n$, it follows that $\sigma(n-1) = n-1$. Working backwards, we see that $\sigma(k) = k$ for all k . Thus, σ is the identity, and all other elementary products are zero. Therefore, $\det(A) = a_{11}a_{22} \cdots a_{nn}$. $\square$

Theorem 2.8.21. *If $A \in M_n(\mathbb{F})$, then $\det(A) = \det(A^T)$.*

Proof. Let $B = [b_{ij}] = A^T$; that is, $b_{ij} = a_{ji}$. It follows that

$$\begin{aligned} \det(B) &= \sum_{\sigma \in S_n} \text{sign}(\sigma) b_{1\sigma(1)} b_{2\sigma(2)} \cdots b_{n\sigma(n)} \\ &= \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}. \end{aligned}$$

Rearranging the elements in each product, we get

$$a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} = a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} \cdots a_{n\sigma^{-1}(n)}.$$

As σ runs through all the elements of S_n , so does σ^{-1} ; thus, by writing $\tau = \sigma^{-1}$, we have

$$\begin{aligned} \det(B) &= \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} \\ &= \sum_{\tau \in S_n} \text{sign}(\tau) a_{1\tau(1)} a_{2\tau(2)} \cdots a_{n\tau(n)} = \det(A). \quad \square \end{aligned}$$

2.9 Determinants II

In this section we show that determinants can be computed using row reduction. This provides a practical approach to computing the determinant, as well as a way to prove several important properties of determinants. We also show that the determinant can be computed using cofactor expansions. From there we prove Cramer's rule and define the *adjugate* of a matrix, which relates the determinant to the inverse of a matrix.

2.9.1 Row Operations on Determinants

Theorem 2.9.1. *If $A, B \in M_n(\mathbb{F})$ and if E is an elementary matrix such that $B = EA$, then $\det(B) = \det(E)\det(A)$. Specifically, if E is of*

- (i) *type I* ($R_k \leftrightarrow R_\ell$), then $\det(B) = -\det(A)$;
- (ii) *type II* ($R_k \rightarrow \alpha R_k$), then $\det(B) = \alpha \det(A)$ (this holds even if $\alpha = 0$);
- (iii) *type III* ($R_\ell \rightarrow \alpha R_k + R_\ell$), with $k \neq \ell$, then $\det(B) = \det(A)$.

Proof. Let $A = [a_{ij}]$ and $B = [b_{ij}]$.

- (i) Let τ be the transposition $k \leftrightarrow l$ (assume $k < l$). Note that $b_{ij} = a_{\tau(i)j}$ and that $\tau^{-1} = \tau$. Thus,

$$\begin{aligned}
 \det(B) &= \sum_{\sigma \in S_n} \text{sign}(\sigma) b_{1\sigma(1)} b_{2\sigma(2)} \cdots b_{k\sigma(k)} \cdots b_{l\sigma(l)} \cdots b_{n\sigma(n)} \\
 &= \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{l\sigma(k)} \cdots a_{k\sigma(l)} \cdots a_{n\sigma(n)} \\
 &= \sum_{\sigma \in S_n} \text{sign}(\tau^{-1}) \text{sign}(\sigma\tau) a_{1\sigma(\tau(1))} a_{2\sigma(\tau(2))} \cdots a_{n\sigma(\tau(n))} \\
 &= - \sum_{\nu \in S_n} \text{sign}(\nu) a_{1\nu(1)} a_{2\nu(2)} \cdots a_{n\nu(n)} = -\det(A).
 \end{aligned}$$

Here the last equality comes from setting $\nu = \sigma\tau$ and noticing that when σ ranges through every element of S_n , then so does ν .

- (ii) Note that $b_{ij} = a_{ij}$ when $i \neq k$ and $b_{ij} = \alpha a_{ij}$ when $i = k$. Thus,

$$\begin{aligned}
 \det(B) &= \sum_{\sigma \in S_n} \text{sign}(\sigma) b_{1\sigma(1)} \cdots b_{n\sigma(n)} \\
 &= \sum_{\sigma \in S_n} \alpha \text{sign}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)} = \alpha \det(A).
 \end{aligned}$$

- (iii) Note that $b_{ij} = a_{ij}$ when $i \neq l$ and $b_{ij} = a_{lj} + \alpha a_{kj}$ when $i = l$. Thus,

$$\begin{aligned}
 \det(B) &= \sum_{\sigma \in S_n} \text{sign}(\sigma) b_{1\sigma(1)} b_{2\sigma(2)} \cdots b_{n\sigma(n)} \\
 &= \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{1\sigma(1)} \cdots (a_{l\sigma(l)} + \alpha a_{k\sigma(l)}) \cdots a_{n\sigma(n)} \\
 &= \det(A) + \alpha \det(C),
 \end{aligned}$$

where $C = [c_{ij}]$ satisfies $c_{ij} = a_{ij}$ when $i \neq l$ and $c_{ij} = a_{kj}$ when $i = l$. Since two of the rows of C are identical, interchanging those two rows cannot change the determinant of C , but we know from part (i) that interchanging any two rows changes the sign of the $\det(C)$. Therefore, $\det(C) = -\det(C)$, which can only be true if $\det(C) = 0$, and so we have $\det(B) = \det(A)$. $\square$

In the proof of (iii) above, we also proved the following corollary of (i).

Corollary 2.9.2. *If $A \in M_n(\mathbb{F})$ has two identical rows, then $\det(A) = 0$.*

Remark 2.9.3. For notational convenience, we often write the determinant of the matrix $A = [a_{ij}] \in M_n(\mathbb{F})$ as

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

We can now perform row operations to compute the determinant.

Example 2.9.4. Using row reduction, we compute the determinant of the matrix

$$A = \begin{bmatrix} -2 & -4 & -6 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}.$$

We perform the following operations:

$$\begin{aligned} \begin{vmatrix} -2 & -4 & -6 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} &= -2 \cdot \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = -2 \cdot \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{vmatrix} \\ &= 6 \cdot \begin{vmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & -6 & -11 \end{vmatrix} = 6 \cdot \begin{vmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix} = 6. \end{aligned}$$

Corollary 2.9.5. If $A \in M_n(\mathbb{F})$, then $\det(\alpha A) = \alpha^n \det(A)$, $\alpha \in \mathbb{F}$.

Proof. The proof is Exercise 2.49. $\square$

Corollary 2.9.6. A matrix $A \in M_n(\mathbb{F})$ is nonsingular if and only if $\det(A) \neq 0$.

Proof. The matrix A is nonsingular if and only if it is row equivalent to the identity (see Theorem 2.7.14). But by the previous theorem, row equivalence changes the determinant only by a nonzero scalar (recall that we do not allow $\alpha = 0$ in the type II elementary matrices). Thus, if A is invertible, then its determinant is a nonzero scalar times $\det(I) = 1$, so $\det(A)$ is nonzero. On the other hand, if A is singular, then it row reduces to an upper-triangular matrix with at least one zero on the diagonal, and hence (by Theorem 2.8.20) it has determinant zero. $\square$

Theorem 2.9.7. If $A, B \in M_n(\mathbb{F})$, then $\det(AB) = \det(A) \det(B)$.

Proof. If A is singular, then AB is also singular by Exercise 2.11. Thus, both $\det(AB) = 0$ and $\det(A) = 0$ by Corollary 2.9.6. If A is nonsingular, then A can

be written as a product of elementary matrices $A = E_k E_{k-1} \cdots E_1$. It follows that

$$\begin{aligned} \det(AB) &= \det(E_k E_{k-1} \cdots E_1 B) = \det(E_k) \det(E_{k-1}) \cdots \det(E_1) \det(B) \\ &= \det(E_k E_{k-1} \cdots E_1) \det(B) = \det(A) \det(B). \quad \square \end{aligned}$$

Corollary 2.9.8. *If $A \in M_n(\mathbb{F})$ is nonsingular, then $\det(A^{-1}) = 1/\det(A)$.*

Proof. Since $1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1})$, the desired result follows. $\square$

Corollary 2.9.9. *If $A, B \in M_n(\mathbb{F})$ are similar matrices, that is, $B = P^{-1}AP$ for some invertible matrix $P \in M_n(\mathbb{F})$, then $\det(A) = \det(B)$.*

Proof. $\det(B) = \det(P^{-1}AP) = \det(P^{-1}) \det(A) \det(P) = \det(A)$. $\square$

Remark 2.9.10. Since the determinant is preserved under similarity transformations, we can define the determinant of a linear operator L on a finite-dimensional vector space as the determinant of any of its matrix representations.

2.9.2 Cofactor Expansions and the Adjugate

The cofactor expansion gives another approach for defining the determinant of a matrix. We use this approach to construct the *adjugate*, which can be used to give an explicit formula for the inverse of a nonsingular matrix. Although the formula is not well suited to numerical computation, it does give us some powerful theoretical tools, most notably Cramer's rule.

Definition 2.9.11. *The (i, j) submatrix A_{ij} of $A \in M_n(\mathbb{F})$ is the $(n-1) \times (n-1)$ matrix obtained by removing row i and column j from A . The (i, j) minor M_{ij} of A is the determinant of A_{ij} . We call $(-1)^{i+j} M_{ij}$ the (i, j) cofactor of A .*

Example 2.9.12. Using the matrix A from Example 2.9.4, we have that

$$A_{23} = \begin{bmatrix} -2 & -4 \\ 7 & 8 \end{bmatrix}.$$

Moreover, the $(2, 3)$ minor of A is $M_{23} = 12$, whereas the $(2, 3)$ cofactor of A is equal to $(-1)^5(12) = -12$.

Lemma 2.9.13. *If A_{ij} is the (i, j) submatrix of $A \in M_n(\mathbb{F})$, then the signed elementary products associated with $(-1)^{i+j} a_{ij} A_{ij}$ are the same as those associated with A .*

Remark 2.9.14. How can the elementary products associated with $(-1)^{i+j}a_{ij}A_{ij}$ be the same as those associated with A ? Suppose we fix i and j and then sort the elementary products of the matrix into two groups: those that contain a_{ij} , and those that do not. We find that there are $(n-1)!$ elementary products that contain a_{ij} and $(n-1) \cdot (n-1)!$ elementary products that do not contain a_{ij} . So, the number of elementary products of A_{ij} is the same as the number of elementary products of A that contain a_{ij} .

Recall also that an elementary product contains exactly one element from each row of A and each column of A (see Remark 2.8.15). The elementary products of A that contain a_{ij} as a factor cannot have another element that lies in row i or column j as a factor; in fact the other factors of the elementary product must come from precisely the submatrix A_{ij} . All that remains is to verify the correct sign of the elementary product and to formalize the proof of the lemma. This is done below.

Proof. Let $i, j \in \{1, 2, \dots, n\}$ be given. Each $\hat{\sigma} \in S_{n-1}$ corresponds to an elementary product in the minor A_{ij} . We define $\sigma \in S_n$ so that the resulting elementary product of A is the same as that which comes by multiplying a_{ij} by the elementary product of A_{ij} coming from $\hat{\sigma}$. Specifically, we have

$$\sigma(k) = \begin{cases} \hat{\sigma}(k), & k < i \text{ and } \hat{\sigma}(k) < j, \\ \hat{\sigma}(k) + 1, & k < i \text{ and } \hat{\sigma}(k) \geq j, \\ j, & k = i, \\ \hat{\sigma}(k-1), & k > i \text{ and } \hat{\sigma}(k-1) < j, \\ \hat{\sigma}(k-1) + 1, & k > i \text{ and } \hat{\sigma}(k-1) \geq j. \end{cases}$$

The number of inversions of σ equals the number of inversions of $\hat{\sigma}$ plus the number of inversions that involve j . Thus, we need only find the number of inversions that depend on j .

Let k be the number of elements of the n -tuple $\sigma = [\sigma(1), \sigma(2), \dots, \sigma(n)]$ that are greater than j , yet whose position in the n -tuple precede that of j . Thus, $n - j - k$ is the number of elements of σ greater than j that follow j . Since there are $n - i$ spots that follow j , it follows that there are $(n - i) - (n - j - k)$ elements smaller than j that follow j . Hence, there are $k + (n - i) - (n - j - k) = i + j + 2(k - i)$ inversions due to j . Since $(-1)^{i+j+2(k-i)} = (-1)^{i+j}$, the signs match the elementary products of A with those in $(-1)^{i+j}a_{ij}A_{ij}$. $\square$

Example 2.9.15. Let $i = 2$, $j = 5$, and $n = 7$. If $\hat{\sigma} = [4, 3, 2, 1, 6, 5]$, then

$$\begin{aligned} \sigma(1) &= \hat{\sigma}(1) = 4 \text{ since } 1 < 2 \text{ and } \hat{\sigma}(1) < 5, \\ \sigma(2) &= 5 \text{ since } 2 = 2, \\ \sigma(3) &= \hat{\sigma}(3-1) = \hat{\sigma}(2) = 3 \text{ since } 3 > 2 \text{ and } \hat{\sigma}(2) < 5, \\ \sigma(4) &= \hat{\sigma}(4-1) = \hat{\sigma}(3) = 2 \text{ since } 4 > 2 \text{ and } \hat{\sigma}(3) < 5, \\ \sigma(5) &= \hat{\sigma}(5-1) = \hat{\sigma}(4) = 1 \text{ since } 5 > 2 \text{ and } \hat{\sigma}(4) < 5, \end{aligned}$$

$$\begin{aligned}\sigma(6) &= \hat{\sigma}(6-1) + 1 = \hat{\sigma}(5) + 1 = 7 \text{ since } 6 > 2 \text{ and } \hat{\sigma}(5) \geq 5, \\ \sigma(7) &= \hat{\sigma}(7-1) + 1 = \hat{\sigma}(6) + 1 = 6 \text{ since } 7 > 2 \text{ and } \hat{\sigma}(6) \geq 5.\end{aligned}$$

Thus, $\sigma = [4, 5, 3, 2, 1, 7, 6]$.

The next theorem gives an alternative way to compute the determinant of a matrix, called the *cofactor expansion*.

Theorem 2.9.16. *Let M_{ij} be the (i, j) minor of $A \in M_n(\mathbb{F})$. If $A = [a_{ij}]$, then*

$$(i) \text{ for any fixed } j \in \{1, \dots, n\}, \text{ we have } \det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} M_{ij};$$

$$(ii) \text{ for any fixed } i \in \{1, \dots, n\}, \text{ we have } \det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij}.$$

Proof. For each (i, j) pair, there are $(n-1)!$ elementary products in A_{ij} . By taking a column sum of cofactors of A (as in (i)), or by taking a row sum (as in (ii)), we get $n!$ unique elementary products, and by the previous lemma, each of these occurs with the correct sign. $\square$

Example 2.9.17. Theorem 2.9.16 says that we can split the determinant into a sum like the one below,

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \cdot \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \cdot \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \cdot \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix},$$

or alternatively as

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \cdot \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \cdot \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \cdot \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

Example 2.9.18. If

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 1 & 1 & 2 & 3 \\ 0 & 2 & 0 & 0 \end{bmatrix},$$

we may expand $\det(A)$ in cofactors along the bottom row (choose $i = 4$ in

Theorem 2.9.16(ii)) as

$$\det(A) = \sum_{j=1}^4 (-1)^{4+j} a_{4j} M_{4j} = 2M_{42}.$$

Next, we compute M_{42} by expanding along the first column (choose $j = 1$ in Theorem 2.9.16(i)). We find

$$\begin{aligned} M_{42} &= \begin{vmatrix} 1 & 3 & 4 \\ 5 & 7 & 8 \\ 1 & 2 & 3 \end{vmatrix} = \sum_{i=1}^3 (-1)^{i+1} a_{i1} \tilde{M}_{i1} \\ &= 1\tilde{M}_{11} - 5\tilde{M}_{21} + 1\tilde{M}_{31} \\ &= (21 - 16) - 5(9 - 8) + (24 - 28) = -4, \end{aligned}$$

where $\tilde{M}_{ij}$ corresponds to the minors of the 3×3 matrix A_{42} . It follows that $\det(A) = 2(-4) = -8$.

Definition 2.9.19. Assume M_{ij} is the (i, j) minor of $A = [a_{ij}] \in M_n(\mathbb{F})$. The adjugate $\text{adj}(A) = [c_{ij}]$ of A is given by $c_{ij} = (-1)^{i+j} M_{ji}$ (note that the indices of the cofactors are transposed $(i, j) \leftrightarrow (j, i)$ in the adjugate).

Remark 2.9.20. Some people call this matrix the *classical adjoint*, but this is a confusing name because the adjoint has another meaning that is more standard (used in Chapter 3). We always call the matrix in Definition 2.9.19 the *adjugate*.

Example 2.9.21. Consider the matrix A from Example 2.9.4. Exercise 2.48 shows that

$$\text{adj}(A) = \begin{bmatrix} 2 & -8 & 6 \\ 2 & 22 & -12 \\ -3 & -12 & 6 \end{bmatrix}.$$

The adjugate can be used to give an explicit formula for the inverse of a matrix. Although this formula is not well suited to computation, it is a powerful theoretical tool.

Theorem 2.9.22. If $A \in M_n(\mathbb{F})$, then $A \text{adj}(A) = \det(A)I$.

Proof. Let $A = [a_{ij}]$ and $\text{adj}(A) = [c_{ij}]$. Note that the (i, k) entry of $A \text{adj}(A)$ is

$$\sum_{j=1}^n a_{ij} c_{jk} = \sum_{j=1}^n (-1)^{k+j} a_{ij} M_{kj}.$$

Now if $i = k$, then the (i, i) entry of $A \text{adj}(A)$ is $\sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij}$, which is equal to $\det(A)$ by Theorem 2.9.16.

Consider now the matrix B which is created by replacing row k of matrix A by row i of matrix A . By Theorem 2.9.16,

$$\det(B) = \sum_{j=1}^n (-1)^{k+j} b_{kj} B_{kj} = \sum_{j=1}^n (-1)^{k+j} b_{ij} B_{kj} = \sum_{j=1}^n (-1)^{k+j} a_{ij} M_{kj}.$$

However, by Corollary 2.9.2, $\det(B) = 0$. Applying this to the theorem at hand, if $i \neq k$, then the (i, k) entry of $A \operatorname{adj}(A)$ is $\sum_{j=1}^n (-1)^{k+j} a_{ij} M_{kj} = 0$. Thus,

$$\sum_{j=1}^n a_{ij} c_{jk} = \sum_{j=1}^n (-1)^{k+j} a_{ij} M_{kj} = \begin{cases} \det(A) & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases} \quad \square$$

Corollary 2.9.23. *If $A \in M_n(\mathbb{F})$ is nonsingular, then $A^{-1} = \frac{\operatorname{adj}(A)}{\det(A)}$.*

One of the most important applications of the adjugate is Cramer's rule, which allows us to write an explicit formula for the solution to a system of linear equations. Although the equations become unwieldy very rapidly, they are useful for proving various properties of the solutions, and this can often make solving the system much simpler.

Corollary 2.9.24 (Cramer's Rule). *If $A \in M_n(\mathbb{F})$ is nonsingular, then the unique solution to $A\mathbf{x} = \mathbf{b}$ is*

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{\operatorname{adj}(A)}{\det(A)}\mathbf{b}. \quad (2.15)$$

Moreover, if $A_i(\mathbf{b}) \in M_n(\mathbb{F})$ is the matrix A with the i th column replaced by $\mathbf{b}$, then the i th coordinate of $\mathbf{x}$ is

$$x_i = \frac{\det(A_i(\mathbf{b}))}{\det(A)}. \quad (2.16)$$

Proof. Note that (2.15) is an immediate consequence of Corollary 2.9.23. To prove (2.16), let $A = [a_{ij}]$, $\mathbf{b} = [b_1 \ b_2 \ \dots \ b_n]^\top$, and $\operatorname{adj}(A) = [c_{ij}]$. Thus,

$$x_i = \frac{\sum_{j=1}^n c_{ij} b_j}{\det(A)} = \sum_{j=1}^n \frac{(-1)^{i+j} b_j M_{ji}}{\det(A)} = \frac{\det(A_i(\mathbf{b}))}{\det(A)}. \quad \square$$

Example 2.9.25. Consider the linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} -2 & -4 & -6 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix},$$

as in Example 2.9.4, and $\mathbf{b} = [6 \ 6 \ 1]^T$. Using (2.16), we compute

$$\det(A_1(\mathbf{b})) = \begin{vmatrix} 6 & -4 & -6 \\ 6 & 5 & 6 \\ 1 & 8 & 10 \end{vmatrix} = -30, \quad \det(A_2(\mathbf{b})) = \begin{vmatrix} -2 & 6 & -6 \\ 4 & 6 & 6 \\ 7 & 1 & 10 \end{vmatrix} = 132,$$

$$\text{and } \det(A_3(\mathbf{b})) = \begin{vmatrix} -2 & -4 & 6 \\ 4 & 5 & 6 \\ 7 & 8 & 1 \end{vmatrix} = -84.$$

Since $\det A = 6$, we have that $\mathbf{x} = [-5 \ 22 \ -14]^T$.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 2.1. Determine whether each of the following is a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$. If it is a linear transformation, give $\mathcal{N}(L)$ and $\mathcal{R}(L)$.
 - (i) $L(x, y) = (y, x)$.
 - (ii) $L(x, y) = (x, 0)$.
 - (iii) $L(x, y) = (x + 1, y + 1)$.
 - (iv) $L(x, y) = (x^2, y^2)$.
- 2.2. Recall the vector space $V = (0, \infty)$ given in Exercise 1.1. Show that the function $T(x) = \log x$ is a linear transformation from V into $\mathbb{R}$.
- 2.3. Determine whether each of the following is a linear transformation from $\mathbb{F}[x; 2]$ into $\mathbb{F}[x; 4]$. Prove your answers.
 - (i) L maps any polynomial $p(x)$ to x^2 ; that is, $L[p](x) = x^2$.
 - (ii) L maps any polynomial $p(x)$ to $xp(x)$; that is, $L[p](x) = xp(x)$.

(iii) $L[p](x) = x^4 + p(x)$.

(iv) $L[p](x) = (4x^2 - 3x)p'(x)$.

- 2.4. Let $L : C^1([0, 1]; \mathbb{F}) \rightarrow C([0, 1]; \mathbb{F})$ be given by $L[f] = f' + f$. Show that L is linear. For $0 \leq x \leq 1$, define

$$f(x) = e^{-x} \int_0^x g(t)e^t dt + Ce^{-x}.$$

Verify that $L[f] = g$.

- 2.5. Prove Corollary 2.2.3.

- 2.6. Let $\{V_i\}_{i=1}^{n+1}$ be a collection of vector spaces and $\{L_i\}_{i=1}^n$ a collection of invertible linear maps $L_i : V_i \rightarrow V_{i+1}$. Prove that $(L_n L_{n-1} \cdots L_1)^{-1} = L_1^{-1} L_2^{-1} \cdots L_n^{-1}$.

- 2.7. Prove Proposition 2.2.14(iv).

- 2.8. Let L be an operator on the vector space V and $k \in \mathbb{N}$. Prove that

(i) $\mathcal{N}(L^k) \subset \mathcal{N}(L^{k+1})$,

(ii) $\mathcal{R}(L^{k+1}) \subset \mathcal{R}(L^k)$.

Note: We can assume that L^0 is the identity operator.

- 2.9. Elaborating on Example 2.1.3(xii), consider the rotation matrix

$$R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}. \quad (2.17)$$

Show that multiplication of column vectors $\begin{bmatrix} x & y \end{bmatrix}^\top \in \mathbb{F}^2$ by R_θ rotates vectors counterclockwise. This is the same as $\rho_\theta(x, y)$ in the example, but it is written as matrix-vector multiplication.

- 2.10. Show that the rotation matrix in the previous exercise satisfies $R_{\theta+\phi} = R_\theta R_\phi$.

- 2.11. Let L_1 and L_2 be operators on a finite-dimensional vector space. Prove that if L_1 is not invertible; then the product $L_1 L_2$ is not invertible.

- 2.12.† Let L be a linear operator on a finite-dimensional vector space. For $k \in \mathbb{N}$, prove the following:

(i) If $\mathcal{N}(L^k) = \mathcal{N}(L^{k+1})$, then $\mathcal{N}(L^\ell) = \mathcal{N}(L^k)$ for all $\ell \geq k$.

(ii) If $\mathcal{R}(L^{k+1}) = \mathcal{R}(L^k)$, then $\mathcal{R}(L^\ell) = \mathcal{R}(L^k)$ for all $\ell \geq k$.

(iii) If $\mathcal{R}(L^{k+1}) = \mathcal{R}(L^k)$, then $\mathcal{R}(L^k) \cap \mathcal{N}(L^k) = \{0\}$.

- 2.13. Let V, W, X be finite-dimensional vector spaces and let $L : V \rightarrow W$ and $K : W \rightarrow X$ be linear transformations. Prove that

$$\text{rank}(KL) = \text{rank}(L) - \dim(\mathcal{N}(K) \cap \mathcal{R}(L)).$$

Hint: Use the rank-nullity theorem (Corollary 2.3.9) on $K_{\mathcal{R}(L)} : \mathcal{R}(L) \rightarrow X$, which is K restricted to the domain $\mathcal{R}(L)$.

2.14. Given the setup of Exercise 2.13, prove the following inequalities:

- (i) $\text{rank}(KL) \leq \min(\text{rank}(L), \text{rank}(K))$.
- (ii) $\text{rank}(K) + \text{rank}(L) - \dim(W) \leq \text{rank}(KL)$.

2.15. Let $\{W_1, W_2, \dots, W_n\}$ be a collection of subspaces of the vector space V . Show that the mapping $L : V \rightarrow V/W_1 \times V/W_2 \times \dots \times V/W_n$ defined by $L(\mathbf{v}) = (\mathbf{v} + W_1, \mathbf{v} + W_2, \dots, \mathbf{v} + W_n)$ is a linear transformation and $\mathcal{N}(L) = \bigcap_{i=1}^n W_i$. This is the vector-space analogue of the *Chinese remainder theorem*. The Chinese remainder theorem is treated in more detail in Chapter 15.

2.16.* Let V be a vector space and suppose that $S \subseteq T \subseteq V$ are subspaces of V . Prove

$$\frac{V/S}{T/S} \cong \frac{V}{T}.$$

2.17. Let L be a linear operator on $\mathbb{F}^2$ that maps the basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ as follows:

$$L(\mathbf{e}_1) = \mathbf{e}_1 + 2\mathbf{e}_2 \quad \text{and} \quad L(\mathbf{e}_2) = 2\mathbf{e}_1 - \mathbf{e}_2.$$

- (i) Compute $L(2\mathbf{e}_1 - 3\mathbf{e}_2)$ and $L^2(2\mathbf{e}_1 - 3\mathbf{e}_2)$ in terms of $\mathbf{e}_1$ and $\mathbf{e}_2$.
 - (ii) Determine the matrix representations of L and L^2 .
- 2.18. Assuming the polynomial bases $[1, x, x^2]$ and $[1, x, x^2, x^3, x^4]$ for $\mathbb{F}[x; 2]$ and $\mathbb{F}[x; 4]$, respectively, find the matrix representations for each of the linear transformations in Exercise 2.3.
- 2.19. Let $L : \mathbb{F}[x; 2] \rightarrow \mathbb{F}[x; 2]$ be given by $L[p] = p + 4p'$. Find the matrix representation of L with respect to the basis $S = [x^2 + 1, x - 1, 1]$ (for both the domain and codomain).
- 2.20. Let $\alpha \neq 0$ be fixed, and let V be the space of infinitely differentiable real-valued functions spanned by $S = [e^{\alpha x}, xe^{\alpha x}, x^2e^{\alpha x}]$. Let D be a linear operator on V given by $D[f](x) = f'$. Find the matrix A representing D on S . (Tip: Save your answer. You need it to solve Exercise 2.41.)
- 2.21. Recall Euler's identity: $e^{i\theta} = \cos \theta + i \sin \theta$. Consider $C([0, 2\pi]; \mathbb{C})$ as a vector space over $\mathbb{C}$. Let V be the subspace spanned by the vectors $S = [e^{i\theta}, e^{-i\theta}]$. Given that $T = [\cos \theta, \sin \theta]$ is also a basis for V ,
- (i) find the transition matrix from S into T ;
 - (ii) find the transition matrix from T into S ;
 - (iii) verify that the transition matrices are inverses of each other.
- 2.22. Let V be the vector space V of the previous exercise. Let $D : V \rightarrow V$ be the derivative operator $D(f) = \frac{d}{d\theta} f(\theta)$. Write the matrix representation of D in terms of
- (i) the basis S on both domain and codomain;
 - (ii) the basis T on both domain and codomain;
 - (iii) the basis S on the domain and basis T on the codomain.
- 2.23. Given a linear transformation defined by a matrix A , prove that the range of the transformation is the span of the columns of A .

-
- 2.24. Show that the matrix representation of the operator $L_2 : \mathbb{F}[x; 4] \rightarrow \mathbb{F}[x; 4]$ given by $L_2[p](x) = p''(x)$ on the standard polynomial basis $[1, x, x^2, x^3, x^4]$ equals $(C_{SS})^2$, as given in Example 2.4.8. In other words, the matrix representation of the second derivative operator is the square of the matrix representation of the derivative operator.
- 2.25. The trace of an $n \times n$ matrix $A = [a_{ij}]$, denoted $\text{tr}(A)$, is the sum of its diagonal entries; that is, $\text{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$. Show the following for any $n \times n$ matrices $A, B \in M_n(\mathbb{R})$ (where $B = [b_{ij}]$):
- (i) $\text{tr}(AB) = \text{tr}(BA)$.
 - (ii) If A and B are similar, then $\text{tr}(A) = \text{tr}(B)$. This implies, among other things, that the trace is determined only by the linear operator and does not depend on the choice of basis we use to write the matrix representation of that operator.
 - (iii) $\text{tr}(AB^T) = \sum_{i=1}^n \sum_{j=1}^n a_{ij}b_{ij}$.
- 2.26. Let $p(z) = a_k z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0$ be a polynomial. For $A \in M_n(\mathbb{F})$, we define $p(A) = a_k A^k + a_{k-1} A^{k-1} + \cdots + a_1 A + a_0 I$. Prove: If A and B are similar, then $p(A)$ and $p(B)$ are also similar.
- 2.27. Prove that similarity is an equivalence relation on the set of all $n \times n$ matrices. What are all the equivalence classes of 1×1 matrices?
- 2.28. Prove that if two matrices are similar and one is invertible, then so is the other.
- 2.29. Prove that similarity preserves rank and nullity of matrices.
- 2.30. Prove: If $\mathbb{F}^m \cong \mathbb{F}^n$, then $m = n$. Hint: Consider using the rank-nullity theorem.
-

- 2.31. A higher-degree Bernstein polynomial can be written in terms of lower-degree Bernstein polynomials. Let $B_n^{n-1}(x) \equiv 0$ and $B_{-1}^{n-1}(x) \equiv 0$ for all $n \geq 0$. Show that for any $n \geq 1$ we have

$$B_j^n(x) = (1-x)B_j^{n-1}(x) + xB_{j-1}^{n-1}(x), \quad j = 0, 1, \dots, n.$$

- 2.32. Since the obvious inclusion $i_{m,n} : \mathbb{F}[x; m] \rightarrow \mathbb{F}[x; n]$ is a linear transformation for every $0 \leq m \leq n$, and since the Bernstein polynomials form a basis of $\mathbb{F}[x; n]$, any lower-degree Bernstein polynomial can be written as a linear combination of higher-degree Bernstein polynomials. Show that

$$B_j^n(x) = \frac{j+1}{n+1} B_{j+1}^{n+1}(x) + \frac{n-j+1}{n+1} B_j^{n+1}(x), \quad j = 0, 1, \dots, n.$$

Write the matrix representation of $i_{n,n+1} : \mathbb{F}[x; n] \rightarrow \mathbb{F}[x; n+1]$ in terms of the Bernstein bases.

- 2.33. The derivative gives a linear transformation $D : \mathbb{F}[x; n] \rightarrow \mathbb{F}[x; n-1]$. Since the set of Bernstein polynomials $B = \{B_0^{n-1}(x), B_1^{n-1}(x), B_2^{n-1}(x), \dots, B_{n-1}^{n-1}(x)\}$ forms a basis for $\mathbb{F}[x; n-1]$, one can always write the derivative of $B_j^n(x)$ as a linear combination of the Bernstein polynomial basis B .

- (i) Show that the derivative of $B_j^n(x)$ can be written as

$$D[B_j^n](x) = n(B_{j-1}^{n-1}(x) - B_j^{n-1}(x)).$$

- (ii) Write the matrix representation of the derivative $D : \mathbb{F}[x; n] \rightarrow \mathbb{F}[x; n-1]$ in terms of the Bernstein bases.
- (iii) The previous step is different from Example 2.6.6 because the domain and codomain of the derivative are different. Use the previous step and Exercise 2.32 to write a product of matrices that gives the matrix representation of the derivative $\mathbb{F}[x; n] \rightarrow \mathbb{F}[x; n]$ (as an operator) in terms of the Bernstein basis.
- 2.34. The map $I : \mathbb{F}[x; n] \rightarrow \mathbb{F}$ given by $f \mapsto \int_0^1 f(x) dx$ is a linear transformation. Write the matrix representation of this linear transformation in terms of the Bernstein basis for $\mathbb{F}[x; n]$ and the (single-element) standard basis for $\mathbb{F}$.
- 2.35.† Define the n th Bernstein operator $B_n : C([0, 1]; \mathbb{F}) \rightarrow \mathbb{F}[x; n]$ to be the linear transformation

$$B_n[f](x) = \sum_{j=0}^n f(j/n) B_j^n(x).$$

For any subspace $W \subset C([0, 1]; \mathbb{F})$, the transformation B_n also defines a linear transformation $W \rightarrow \mathbb{F}[x; n]$.

- (i) Find the matrix representation of $B_1 : \mathbb{F}[x; 2] \rightarrow \mathbb{F}[x; 1]$ in terms of the Bernstein bases.
- (ii) Let $V \subset C([0, 1]; \mathbb{F})$ be the two-dimensional subspace spanned by the set $S = [\sin x, \cos x]$. For every $n \geq 0$, write the matrix representation of the linear transformation $B_n : V \rightarrow \mathbb{F}[x; n]$ in terms of the basis S of V and the Bernstein basis of $\mathbb{F}[x; n]$.

- 2.36. Using only type III elementary matrices, reduce the matrix

$$A = \begin{bmatrix} 2 & 4 & 6 \\ 4 & 5 & 6 \\ 7 & 8 & 0 \end{bmatrix}$$

to an upper-triangular matrix U . Clearly label each elementary matrix E_i satisfying $U = E_k E_{k-1} \cdots E_1 A$. Then compute $L = (E_k E_{k-1} \cdots E_1)^{-1}$ and verify that $A = LU$.

- 2.37. Reduce each of the following matrices to RREF. Describe the kernel $\mathcal{N}(A)$ of the linear transformation defined by the matrix A . Let $\mathbf{b} = [1 \ 1 \ 0]^T$. Determine whether the system $A\mathbf{x} = \mathbf{b}$ has a solution, and if it does, describe the set of all solutions as a coset of $\mathcal{N}(A)$.

(i) $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$

(ii) $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$

$$(iii) \ A = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 5 & 4 \\ 7 & 7 & 7 \end{bmatrix}.$$

$$(iv) \ A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

$$(v) \ A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 25 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

2.38. 

(i) Let $\mathbf{e}_i$ and $\mathbf{e}_j$ be the i th and j th standard basis elements, respectively, (thought of as column vectors). Prove that the product $\mathbf{e}_i \mathbf{e}_j^T$ is the matrix with a one in its (i, j) entry and zeros everywhere else.

(ii) Let $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$, $a \in \mathbb{F}$, and $a\mathbf{v}^T\mathbf{u} \neq 1$. Prove that

$$(I - a\mathbf{u}\mathbf{v}^T)^{-1} = I - \frac{a\mathbf{u}\mathbf{v}^T}{a\mathbf{v}^T\mathbf{u} - 1}.$$

In particular, $(I - a\mathbf{u}\mathbf{v}^T)$ is nonsingular whenever $a\mathbf{v}^T\mathbf{u} \neq 1$. Hint: Multiply both sides by $(I - a\mathbf{u}\mathbf{v}^T)$ and recall $\mathbf{v}^T\mathbf{u}$ is a scalar.

All three types of elementary matrices are of the form $I - a\mathbf{u}\mathbf{v}^T$. Specifically, type I matrices can be written as $I - (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$, type II matrices can be written as $I - (1 - \alpha)\mathbf{e}_i\mathbf{e}_i^T$, and type III matrices can be written as $I + \alpha\mathbf{e}_i\mathbf{e}_j^T$.

2.39. Prove Proposition 2.7.13. Hint: Do row reduction with elementary matrices.

2.40. Prove that row equivalence is an equivalence relation (see Theorem 2.7.7). Hint: If E is an elementary matrix, then so is E^{-1} .

2.41. Let A be the matrix in Exercise 2.20. Use its inverse to find the antiderivative of $f(x) = ae^x + bxe^x + cx^2e^x$ in the space V . Antiderivatives are not usually unique. Explain why there is a unique solution in this case.

2.42. Let $L : \mathbb{R}[x; 3] \rightarrow \mathbb{R}[x; 3]$ be given by $L[f](x) = (x - 1)f'(x)$. Find bases for both $\mathcal{N}(L)$ and $\mathcal{R}(L)$.

2.43. How many inversions are there in the permutation $\sigma = [1, 4, 3, 2, 6, 5, 9, 8, 7]$?

2.44. List all the permutations in S_4 .

2.45. Use the definition of determinant to compute the determinant of the matrix

$$A = \begin{bmatrix} 0 & 2 & 3 & 0 \\ 4 & 0 & 0 & 5 \\ 0 & 6 & 7 & 1 \\ 8 & 0 & 9 & 0 \end{bmatrix}.$$

2.46. Prove that if a matrix A has a row (or column) of zeros, then $\det(A) = 0$.

- 2.47. Recall (see Definition C.1.3) that the *Hermitian conjugate* A^H of any $m \times n$ matrix is the conjugate transpose

$$A^H = \overline{A}^T = [\overline{a_{ji}}].$$

Prove that for any $A \in M_n(\mathbb{C})$ we have $\det(A^H) = \overline{\det(A)}$.

- 2.48. Compute the adjugate for the matrix in Example 2.9.4.
 2.49. Prove Corollary 2.9.5.
 2.50. Prove: If a matrix can be written in block-triangular form

$$M = \begin{bmatrix} A & B \\ 0 & D \end{bmatrix},$$

then $\det(M) = \det(A) \det(D)$. Hint: Use row operations.

- 2.51. Let $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, and assume $\mathbf{y}^H \mathbf{x} \neq 1$. Show that $\det(I - \mathbf{x} \mathbf{y}^H) = 1 - \mathbf{y}^H \mathbf{x}$.
 Hint: Note that

$$\begin{bmatrix} I & 0 \\ -\mathbf{y}^H & 1 \end{bmatrix} \begin{bmatrix} I - \mathbf{x} \mathbf{y}^H & \mathbf{x} \\ 0^H & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ \mathbf{y}^H & 1 \end{bmatrix} = \begin{bmatrix} I & \mathbf{x} \\ 0^H & 1 - \mathbf{y}^H \mathbf{x} \end{bmatrix}.$$

- 2.52. Prove: If a matrix can be written in block-triangular form

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A is invertible, then $\det(M) = \det(A) \det(D - CA^{-1}B)$. Hint: Use Exercise 2.50 and the fact that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & I \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{bmatrix}.$$

- 2.53. Use Corollary 2.9.23 to compute the inverse of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 5 & 6 & 7 \end{bmatrix}.$$

- 2.54. Use Cramer's rule to solve the system $A\mathbf{x} = \mathbf{b}$ in Exercise 2.37(v).

- 2.55.† Consider the matrix

$$V_n = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}.$$

Prove that

$$\det(V_n) = \prod_{i < j} (x_j - x_i).$$

Hint: Row reduce the transpose. Subtract x_0 times row $k-1$ from row k for $k = 1, \dots, n$. Then factor out all the $(x_k - x_0)$ terms to reduce the problem and proceed recursively.

2.56.† The *Wronskian* gives a sufficient condition that determines whether a set of n functions in the space $C^{n-1}([a, b]; \mathbb{F})$ is linearly independent. The Wronskian $W(x)$ of a set of functions $\mathcal{C} = \{y_1(x), y_2(x), \dots, y_n(x)\}$ is defined to be

$$W(x) = \det \begin{bmatrix} y_1(x) & y_2(x) & \dots & y_n(x) \\ y_1'(x) & y_2'(x) & \dots & y_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)}(x) & y_2^{(n-1)}(x) & \dots & y_n^{(n-1)}(x) \end{bmatrix}, \quad (2.18)$$

where $y_j^{(k)}(x)$ is the k th derivative of $y_j(x)$.

- (i) Prove: If there exists $x \in [a, b]$ such that $W(x) \neq 0$, then the set $\mathcal{C}$ is linearly independent in $C^n([a, b]; \mathbb{F})$. Hint: Prove the contrapositive.
- (ii) Use the Wronskian to determine whether the set S in Exercise 2.20 is linearly independent.

Notes

As mentioned at the end of the previous chapter, some standard references for the material of this chapter include [Lay02, Leo80, OS06, Str80, Axl15] and the video explanations of G. Strang in [Str10].

3

Inner Product Spaces

Without deviation from the norm, progress is not possible.
—Frank Zappa

In the first two chapters, we focused primarily on the *algebraic*¹⁴ properties of vector spaces and linear transformations (which includes matrices). While we have provided some geometric intuition about key concepts involving linearity, subspaces, spans, cosets, isomorphisms, and determinants, these mathematical ideas and structures have largely been algebraic in nature. In this section, we introduce geometric structure into vector spaces by way of the inner product.

Inner products provide vector spaces with two essential geometric properties, namely, *lengths* and *angles*,¹⁵ and in particular the concept of *orthogonality*—a fancy word for saying that two vectors are at right angles. Here we define inner products and describe their role in characterizing the geometry of linear spaces.

Chapter 1 shows that a basis of a given vector space provides a way to represent each vector uniquely as a linear combination of those basis elements. By ordering the basis vectors, we can also represent a given vector as a list of coordinates for that basis, where each coordinate corresponds to the coefficient of that basis vector in the corresponding linear combination. In this chapter, we go a step further and consider *orthonormal*¹⁶ bases, that is, bases that are also orthonormal sets, where, in addition to providing a unique representation of each vector and a coordinatization of the vector space, the orthonormality allows us to use the inner product to project a given vector along each basis element to determine its coefficients. This feature is particularly nice when only some of the coordinates are needed, as is common in many applications, and is in stark contrast to the nonorthonormal case, where

¹⁴By the word *algebraic* we mean anything related to the study of mathematical symbols and the rules for manipulating these symbols.

¹⁵In a real vector space, the inner product gives us angles by way of the *law of cosines* (see Section 3.1), but in a complex vector space, the idea of an angle breaks down somewhat and isn't very intuitive. However, for both real and complex vector spaces, orthogonality is well defined by the inner product and is an extremely useful concept in theory, application, and computation.

¹⁶An orthonormal set is one where each vector is orthogonal to every other vector and all of the vectors are of unit length.

one typically has to solve a linear system in order to determine the coordinates; see Example 1.2.14(iii) and Exercise 1.6 for details.

After establishing this and several other important properties of orthonormal bases, we show that any finite linearly independent set can be transformed into an orthonormal one that preserves the span. This algorithm is called the *Gram–Schmidt orthonormalization method*, and it is widely used in both theory and applications. This means that we can always construct an orthonormal basis for a given finite-dimensional vector space or subspace, and then bring to bear the power of the inner product in determining the coefficients of a given vector in this new orthonormal basis.

More generally, the inner product also gives us the ability to compute the lengths of vectors (or distances between vectors by computing the length of their difference). In other words, the inner product induces a *norm* (or length function) on a vector space. Armed with this induced norm, we can examine many ideas from geometry, including perimeters, angles between vectors and subspaces, and the Pythagorean theorem. A norm function also gives us the ability to establish convergence properties for sequences, which in turn enables us to take limits of vector-valued functions; this is at the heart of the theory of calculus on vector spaces.

Because of the far-reaching consequences of norm functions, we take a brief detour from inner products and devote a section to the general properties of norms. In particular, we consider examples of norm functions that are very useful in mathematical analysis but cannot be induced by an inner product. This allows us to expand our way of thinking and understand the key properties of norm functions apart from inner products and Euclidean geometry. Given a linear map from one normed vector space to another, we can look at how these linear maps transform unit vectors. In particular, we look at the maximum distortion obtained by mapping the set of all unit vectors. We show that this quantity is a norm on the set of linear transformations from the domain to the codomain.

We return to properties of the inner product by considering how they affect linear transformations. We introduce the adjoint map and fundamental subspaces. These ideas help us to decompose the domains and codomains of a linear transformation into complementary subspaces that are orthogonal to one another. As a result, we are able to form orthonormal bases for each of these subspaces. One of the biggest applications of this is least squares, which is fundamental to classical statistics. It gives us the ability to fit lines to data and, more generally, to solve many curve-fitting problems where the unknown coefficients can be formulated as linear combinations of certain basis functions.

3.1 Introduction to Inner Products

In this section, we define the inner product on a vector space and describe some of its essential properties. We also provide several examples.

3.1.1 Definitions and Examples

Definition 3.1.1. An inner product on a vector space V is a scalar-valued map $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{F}$ that satisfies the following conditions for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and any $a, b \in \mathbb{F}$:

- (i) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ with equality holding if and only if $\mathbf{x} = \mathbf{0}$ (positivity).
- (ii) $\langle \mathbf{x}, a\mathbf{y} + b\mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{y} \rangle + b\langle \mathbf{x}, \mathbf{z} \rangle$ (linearity).
- (iii) $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$ (conjugate symmetry).

A vector space V together with an inner product $\langle \cdot, \cdot \rangle$ is called an inner product space and is denoted by the pair $(V, \langle \cdot, \cdot \rangle)$.

Remark 3.1.2. If $\mathbb{F} = \mathbb{R}$, then the conjugate symmetry condition given in (iii) simplifies to $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.

Proposition 3.1.3. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and any $a \in \mathbb{F}$, we have

- (i) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$,
- (ii) $\langle a\mathbf{x}, \mathbf{y} \rangle = \bar{a}\langle \mathbf{x}, \mathbf{y} \rangle$.

Proof.

- (i) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \overline{\langle \mathbf{z}, \mathbf{x} + \mathbf{y} \rangle} = \overline{\langle \mathbf{z}, \mathbf{x} \rangle} + \overline{\langle \mathbf{z}, \mathbf{y} \rangle} = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$.
- (ii) $\langle a\mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, a\mathbf{x} \rangle} = \bar{a}\overline{\langle \mathbf{y}, \mathbf{x} \rangle} = \bar{a}\langle \mathbf{x}, \mathbf{y} \rangle$. $\square$

Remark 3.1.4. From the proposition, we see that an inner product on a real vector space is also linear in its first entry; that is, $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$. Since it is linear in both entries, we say that the inner product on a real vector space is bilinear. For complex vector spaces, however, the conjugate symmetry only makes the inner product *half linear* in the first entry since sums can be pulled out of the inner product, but scalars come out conjugated. Thus, the inner product on a complex vector space is called *sesquilinear*,¹⁷ meaning one-and-a-half linear.

Example 3.1.5. In $\mathbb{R}^n$, the standard inner product (or *dot product*) is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n a_i b_i, \quad (3.1)$$

where $\mathbf{x} = [a_1 \ \cdots \ a_n]^\top$ and $\mathbf{y} = [b_1 \ \cdots \ b_n]^\top$. In $\mathbb{C}^n$ the standard inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y} = \sum_{i=1}^n \bar{a}_i b_i. \quad (3.2)$$

¹⁷Some books define the inner product to be linear in the first entry instead of the second, but linearity in the second entry gives a more natural connection to complex inner products.

Example 3.1.6. In $L^2([a, b]; \mathbb{R})$, the standard inner product is

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx, \quad (3.3)$$

whereas in $L^2([a, b]; \mathbb{C})$, it is

$$\langle f, g \rangle = \int_a^b \overline{f(x)}g(x)dx. \quad (3.4)$$

Example 3.1.7. In $M_{m \times n}(\mathbb{F})$, the standard inner product is

$$\langle A, B \rangle = \text{tr}(A^H B), \quad (3.5)$$

This is called the *Frobenius* (or *Hilbert–Schmidt*) inner product.

Remark 3.1.8. In each of the three examples above, we give the real case and the complex case separately to emphasize the importance of the conjugate term in the inner product. Hereafter, we dispense with the distinction and simply write $\mathbb{F}^n$, $L^2([a, b]; \mathbb{F})$, and $M_{m \times n}(\mathbb{F})$, respectively. In each case, we use the complex version of the inner product, since the complex inner product reduces to the real version if the corresponding field is real.

Remark 3.1.9. Usually when denoting the inner product we just write $\langle \cdot, \cdot \rangle$. However, if the theorem or problem we are studying has multiple vector spaces and/or inner products, we distinguish the different inner products with a subscript denoting the space, for example, $\langle \cdot, \cdot \rangle_V$.

Example 3.1.10. On the vector space ℓ^2 over the field $\mathbb{F}$ (see Examples 1.1.6(iv) and 1.1.6(vi)), the standard inner product is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} \overline{a_i} b_i, \quad (3.6)$$

where $\mathbf{x} = (a_1, a_2, \dots)$ and $\mathbf{y} = (b_1, b_2, \dots)$.

The axioms of an inner product are straightforward to verify, but it is not immediately obvious why the inner product should be finite, that is, that the sum should converge. This follows from Hölder's inequality (3.32), which we prove in Section 3.6.

3.1.2 Lengths and Angles

Throughout the remainder of this section, assume that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space over the field $\mathbb{F}$.

Definition 3.1.11. The length of a vector $\mathbf{x} \in V$ induced by the inner product is $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. If $\|\mathbf{x}\| = 1$, we say that $\mathbf{x}$ is a unit vector. The distance between two vectors $\mathbf{x}, \mathbf{y} \in V$ is the length of the difference; that is, $\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Remark 3.1.12. By Definition 3.1.1(i), we know that $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in V$. We call this property *positivity*. Moreover, we know that $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$. We can also show that the length function preserves scale; that is, $\|a\mathbf{x}\| = \sqrt{\langle a\mathbf{x}, a\mathbf{x} \rangle} = \sqrt{|a|^2 \langle \mathbf{x}, \mathbf{x} \rangle} = |a| \|\mathbf{x}\|$. The function $\|\cdot\|$ also has another key property called the triangle inequality; this is examined in Section 3.5.

Remark 3.1.13. Any nonzero vector $\mathbf{x}$ can be normalized to a unit vector by dividing by its length $\|\mathbf{x}\|$. In other words, $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ is a unit vector since

$$\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\| = \left\| \frac{1}{\|\mathbf{x}\|} \mathbf{x} \right\| = \frac{1}{\|\mathbf{x}\|} \|\mathbf{x}\| = 1.$$

Definition 3.1.14. We say that the vector $\mathbf{x} \in V$ is orthogonal to the vector $\mathbf{y} \in V$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Sometimes we denote this by $\mathbf{x} \perp \mathbf{y}$.

Remark 3.1.15. Note that $\langle \mathbf{y}, \mathbf{x} \rangle = 0$ if and only if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. In other words, orthogonality is a symmetric relation between two vectors.

The zero vector $\mathbf{0}$ is orthogonal¹⁸ to every $\mathbf{x} \in V$, since $\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{0} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle$. In the following proposition, we show that the converse is also true.

Proposition 3.1.16. If $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ for all $\mathbf{x} \in V$, then $\mathbf{y} = \mathbf{0}$.

Proof. If $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ holds for all $\mathbf{x} \in V$, then it holds when $\mathbf{x} = \mathbf{y}$. Thus, we have that $0 = \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{y}\|^2$, which implies that $\mathbf{y} = \mathbf{0}$. $\square$

Proposition 3.1.17 (Cauchy–Schwarz Inequality). For all $\mathbf{x}, \mathbf{y} \in V$, we have

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|, \quad (3.7)$$

with equality holding if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

Proof. See Exercise 3.6. $\square$

¹⁸There is no consensus in the literature as to whether the two vectors in the definition should have to be nonzero. As we see in the following section, it doesn't really matter that much since we are usually more interested in orthonormality, which is when the vectors are orthogonal and of unit length.

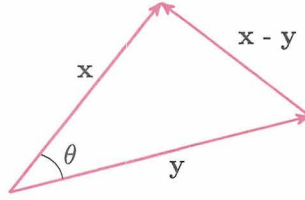


Figure 3.1. Geometric representation of the vectors and angles involved in the law of cosines, as discussed in Remark 3.1.19.

Definition 3.1.18. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space over $\mathbb{R}$. We define the angle between two nonzero vectors $\mathbf{x}$ and $\mathbf{y}$ to be the unique angle $\theta \in [0, \pi]$ such that

$$\cos \theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (3.8)$$

Remark 3.1.19. In $\mathbb{R}^2$ it is straightforward to show that (3.8) follows from the law of cosines (see Figure 3.1), given by

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\| \|\mathbf{y}\| \cos \theta.$$

Note also that this definition does not extend to complex vector spaces. Indeed, the definition of angle breaks down in the complex case because $\langle \mathbf{x}, \mathbf{y} \rangle$ is a complex number.

The Pythagorean theorem is fundamental for classical geometry, but we can now show that it holds for any inner product space, even if the inner product (and hence the length $\|\cdot\|$ of a vector) is very different from what we are used to in classical geometry.

Theorem 3.1.20 (Pythagorean Law). If $\mathbf{x}, \mathbf{y}$ are orthogonal vectors, then $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.

Proof. $\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \|\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2. \quad \square$

Example 3.1.21. In $L^2([0, 1]; \mathbb{C})$ the vectors $f(t) = e^{2\pi i t}$ and $g(t) = e^{10\pi i t}$ are orthogonal, since

$$\langle f, g \rangle = \int_0^1 \overline{e^{2\pi i t}} e^{10\pi i t} dt = \int_0^1 e^{8\pi i t} dt = \frac{e^{8\pi i t}}{8\pi i} \Big|_0^1 = 0.$$

A similar computation shows that $\|f\| = \|g\| = 1$. If $h = 4e^{2\pi i t} + 3e^{10\pi i t}$, then by the Pythagorean law, we have $\|h\|^2 = \|4f\|^2 + \|3g\|^2 = 25$.

3.1.3 Orthogonal Projections

An inner product allows us to define the orthogonal projection of any vector $\mathbf{v}$ onto a unit vector $\mathbf{u}$ (or rather to the one-dimensional subspace spanned by $\mathbf{u}$). This gives us the vector in $\text{span}(\{\mathbf{u}\})$ that is closest (in terms of the norm $\|\cdot\|$) to the original vector $\mathbf{v}$.

Definition 3.1.22. For any vector $\mathbf{x} \in V$, $\mathbf{x} \neq \mathbf{0}$, and any $\mathbf{v} \in V$, the orthogonal projection of $\mathbf{v}$ onto $\text{span}(\{\mathbf{x}\})$ is the vector

$$\text{proj}_{\text{span}(\{\mathbf{x}\})}(\mathbf{v}) = \langle \mathbf{x}, \mathbf{v} \rangle \frac{\mathbf{x}}{\|\mathbf{x}\|^2}. \quad (3.9)$$

Remark 3.1.23. For any nonzero $\alpha \in \mathbb{F}$ we have

$$\text{proj}_{\text{span}(\{\alpha\mathbf{x}\})} = \langle \alpha\mathbf{x}, \mathbf{v} \rangle \frac{\alpha\mathbf{x}}{\|\alpha\mathbf{x}\|^2} = \overline{\alpha}\alpha \langle \mathbf{x}, \mathbf{v} \rangle \frac{\mathbf{x}}{|\alpha|^2\|\mathbf{x}\|^2} = \langle \mathbf{x}, \mathbf{v} \rangle \frac{\mathbf{x}}{\|\mathbf{x}\|^2} = \text{proj}_{\text{span}(\{\mathbf{x}\})},$$

so despite the fact that the definition of $\text{proj}_{\text{span}(\{\mathbf{x}\})}$ depends explicitly on $\mathbf{x}$, it really is determined only by the one-dimensional subspace $\text{span}(\{\mathbf{x}\})$. Nevertheless, for convenience we usually write $\text{proj}_{\mathbf{x}}(\mathbf{v})$, instead of the more cumbersome $\text{proj}_{\text{span}(\{\mathbf{x}\})}(\mathbf{v})$.

Proposition 3.1.24. For any vector $\mathbf{x} \in V$ the map $\text{proj}_{\mathbf{x}} : V \rightarrow V$ is a linear operator. Moreover, the following hold:

- (i) $\text{proj}_{\mathbf{x}} \circ \text{proj}_{\mathbf{x}} = \text{proj}_{\mathbf{x}}$.
- (ii) For any $\mathbf{v} \in V$ the residual vector $\mathbf{r} = \mathbf{v} - \text{proj}_{\mathbf{x}}(\mathbf{v})$ is orthogonal to all the vectors in $\text{span}(\{\mathbf{x}\})$, including $\text{proj}_{\mathbf{x}}(\mathbf{v})$; see Figure 3.2. Thus, $\text{proj}_{\mathbf{x}}(\mathbf{r}) = \mathbf{0}$, or equivalently $\mathbf{r} \in \mathcal{N}(\text{proj}_{\mathbf{x}})$.
- (iii) The vector $\text{proj}_{\mathbf{x}}(\mathbf{v})$ is the unique vector in $\text{span}(\{\mathbf{x}\})$ that is nearest to $\mathbf{v}$. More precisely, $\|\mathbf{v} - \text{proj}_{\mathbf{x}}(\mathbf{v})\| < \|\mathbf{v} - \tilde{\mathbf{x}}\|$ for all $\tilde{\mathbf{x}} \in \text{span}(\{\mathbf{x}\})$ satisfying $\tilde{\mathbf{x}} \neq \text{proj}_{\mathbf{x}}(\mathbf{v})$.

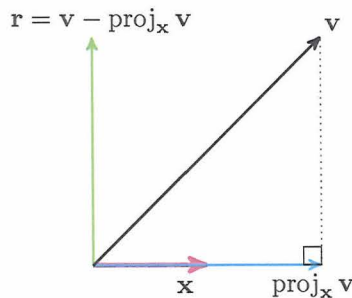


Figure 3.2. Orthogonal projection $\text{proj}_{\mathbf{x}} \mathbf{v} = \langle \mathbf{x}, \mathbf{v} \rangle \mathbf{x}$ (blue) of a vector $\mathbf{v}$ (black) onto a unit vector $\mathbf{x}$ (red). The difference $\mathbf{r} = \mathbf{v} - \text{proj}_{\mathbf{x}} \mathbf{v} = \mathbf{v} - \langle \mathbf{x}, \mathbf{v} \rangle \mathbf{x}$ (green) is orthogonal to $\mathbf{x}$.

Proof. It is clear from the definition that $\text{proj}_{\mathbf{x}}$ is a linear operator. Since the projection depends only on the span of $\mathbf{x}$, we may replace $\mathbf{x}$ by the unit vector $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$

(i) For any $\mathbf{v} \in V$ we have

$$\begin{aligned}\text{proj}_{\mathbf{u}}(\text{proj}_{\mathbf{u}}(\mathbf{v})) &= \langle \mathbf{u}, \text{proj}_{\mathbf{u}}(\mathbf{v}) \rangle \mathbf{u} = \langle \mathbf{u}, \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} \rangle \mathbf{u} \\ &= \langle \mathbf{u}, \mathbf{v} \rangle \langle \mathbf{u}, \mathbf{u} \rangle \mathbf{u} = \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} = \text{proj}_{\mathbf{u}}(\mathbf{v}).\end{aligned}$$

(ii) To show that $a\mathbf{u} \perp \mathbf{r}$ for any $a \in \mathbb{F}$, take the inner product

$$\begin{aligned}\langle a\mathbf{u}, \mathbf{r} \rangle &= \langle a\mathbf{u}, \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}) \rangle = \langle a\mathbf{u}, \mathbf{v} - \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} \rangle \\ &= \langle a\mathbf{u}, \mathbf{v} \rangle - \langle a\mathbf{u}, \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} \rangle = \bar{a} \langle \mathbf{u}, \mathbf{v} \rangle - \bar{a} \langle \mathbf{u}, \mathbf{v} \rangle \langle \mathbf{u}, \mathbf{u} \rangle = 0.\end{aligned}$$

(iii) Given any vector $\tilde{\mathbf{x}} \in \text{span}(\{\mathbf{u}\})$, the square of the distance from $\mathbf{v}$ to $\tilde{\mathbf{x}}$ is

$$\|\mathbf{v} - \tilde{\mathbf{x}}\|^2 = \|\mathbf{r} + \text{proj}_{\mathbf{u}}(\mathbf{v}) - \tilde{\mathbf{x}}\|^2 = \|\mathbf{r}\|^2 + \|\text{proj}_{\mathbf{u}}(\mathbf{v}) - \tilde{\mathbf{x}}\|^2.$$

If $\tilde{\mathbf{x}} \neq \text{proj}_{\mathbf{u}}(\mathbf{v})$, then $\|\text{proj}_{\mathbf{u}}(\mathbf{v}) - \tilde{\mathbf{x}}\|^2 > 0$, and thus $\|\mathbf{v} - \tilde{\mathbf{x}}\|$ is minimized when $\tilde{\mathbf{x}} = \text{proj}_{\mathbf{u}}(\mathbf{v})$. $\square$

Example 3.1.25. Consider the vector space $\mathbb{R}[x]$ with inner product $\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx$. If $f(x) = 5x^3 - 3x$, then

$$\|f\|^2 = \langle f, f \rangle = \int_{-1}^1 (5x^3 - 3x)^2 dx = \frac{8}{7}.$$

The projection of $g(x) = x^3$ onto f is given by

$$\text{proj}_f[g](x) = \left(\int_{-1}^1 (5t^3 - 3t)t^3 dt \right) \frac{7}{8} f(x) = x^3 - \frac{3}{5}x.$$

Moreover, the residual vector is $r(x) = g(x) - \text{proj}_f[g](x) = \frac{3}{5}x$, which by the previous proposition is orthogonal to f . We verify this by a direct check:

$$\langle f, r \rangle = \int_{-1}^1 (5x^3 - 3x) \cdot \frac{3}{5}x dx = 0.$$

3.2 Orthonormal Sets and Orthogonal Projections

The role of orthogonality is essential in many applications. It allows for many problems to be broken up into individual components and solved or analyzed in parts. Orthogonality is also a key concept in a number of computational problems that would be difficult or prohibitive to solve otherwise. In this section, we examine the main properties of orthonormal sets and orthogonal projections.

Throughout this section, assume that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space over the field $\mathbb{F}$.

3.2.1 Properties of Orthonormal Sets

Definition 3.2.1. A collection $\mathcal{C} = \{\mathbf{x}_i\}_{i \in J}$ is an orthonormal set if for all $i, j \in J$ we have

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{ij},$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

is the Kronecker delta.

Example 3.2.2. Let $V = C([0, 1]; \mathbb{R})$. We verify that $S = \{1, \sqrt{12}(x - 1/2)\}$ is an orthonormal set with the inner product

$$\langle f, g \rangle = \int_0^1 f(x)g(x) dx. \quad (3.10)$$

Note that $\langle 1, \sqrt{12}(x - 1/2) \rangle = \sqrt{12} \int_0^1 (x - \frac{1}{2}) dx = \sqrt{12} (\frac{1}{2} - \frac{1}{2}) = 0$. Also $\langle 1, 1 \rangle = \int_0^1 1 dx = 1$ and $\langle \sqrt{12}(x - 1/2), \sqrt{12}(x - 1/2) \rangle = \int_0^1 12(x - 1/2)^2 dx = \int_{-1/2}^{1/2} 12u^2 du = 1$.

Theorem 3.2.3. Let $\{\mathbf{x}_i\}_{i=1}^m$ be an orthonormal set.

- (i) If $\mathbf{x} = \sum_{i=1}^m a_i \mathbf{x}_i$, then $a_i = \langle \mathbf{x}_i, \mathbf{x} \rangle$ for each i .
- (ii) If $\mathbf{x} = \sum_{i=1}^m a_i \mathbf{x}_i$ and $\mathbf{y} = \sum_{i=1}^m b_i \mathbf{x}_i$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^m \bar{a}_i b_i$.
- (iii) If $\mathbf{x} = \sum_{i=1}^m a_i \mathbf{x}_i$, then $\|\mathbf{x}\|^2 = \sum_{i=1}^m |a_i|^2$.

Proof.

- (i) $\langle \mathbf{x}_i, \mathbf{x} \rangle = \langle \mathbf{x}_i, \sum_{j=1}^m a_j \mathbf{x}_j \rangle = \sum_{j=1}^m a_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{j=1}^m a_j \delta_{ij} = a_i$.
- (ii) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \sum_{i=1}^m a_i \mathbf{x}_i, \sum_{j=1}^m b_j \mathbf{x}_j \rangle = \sum_{i=1}^m \sum_{j=1}^m \bar{a}_i b_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{i=1}^m \bar{a}_i b_i$.
- (iii) This follows immediately from (ii). $\square$

Remark 3.2.4. The values $a_i = \langle \mathbf{x}_i, \mathbf{x} \rangle$ in Theorem 3.2.3(i) are called the *Fourier coefficients* of $\mathbf{x}$. The ability to find these coefficients by simply taking the inner product is the hallmark of orthonormal sets.

Corollary 3.2.5. Orthonormal sets are linearly independent. In particular, if $\mathcal{C}$ is an orthonormal set that spans V , then $\mathcal{C}$ is a basis of V .

Proof. Assume that $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m = \mathbf{0}$ for some orthonormal subset $\{\mathbf{x}_i\}_{i=1}^m$ of $\mathcal{C}$. By Theorem 3.2.3(iii), we have that $\sum_{i=1}^m |a_i|^2 = 0$, which implies that each $a_i = 0$. Hence, the set $\mathcal{C}$ is linearly independent. $\square$

We can generalize vector projections from Section 3.1.3 to finite orthonormal sets.

Definition 3.2.6. Let $\{\mathbf{x}_i\}_{i=1}^m$ be an orthonormal set that spans the subspace $X \subset V$. For any $\mathbf{v} \in V$ we define the orthogonal projection onto X to be the sum of vector projections along each $\mathbf{x}_i$; that is,

$$\text{proj}_X(\mathbf{v}) = \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i. \quad (3.11)$$

Theorem 3.2.7. If $\{\mathbf{x}_i\}_{i=1}^m$ is an orthonormal set with $\text{span}(\{\mathbf{x}_i\}_{i=1}^m) = X$, then the map $\text{proj}_X : V \rightarrow V$ is a linear operator. Moreover, the following hold:

- (i) $\text{proj}_X \circ \text{proj}_X = \text{proj}_X$.
- (ii) For every $\mathbf{v} \in V$, the residual vector $\mathbf{r} = \mathbf{v} - \text{proj}_X(\mathbf{v})$ is orthogonal to every $\mathbf{x} \in X$ (see Figure 3.3). Thus, $\text{proj}_X(\mathbf{r}) = \mathbf{0}$, or equivalently $\mathbf{r} \in \mathcal{N}(\text{proj}_X)$.
- (iii) The point $\text{proj}_X(\mathbf{v})$ is the unique vector in X that is nearest to $\mathbf{v}$. More precisely, $\|\mathbf{v} - \text{proj}_X(\mathbf{v})\| < \|\mathbf{v} - \mathbf{x}\|$ for all $\mathbf{x} \in X$ satisfying $\mathbf{x} \neq \text{proj}_X(\mathbf{v})$.

Proof. It is straightforward to check that proj_X is linear.

- (i) For any $\mathbf{v} \in V$ we have (Theorem 3.2.3(i)) that $\langle \mathbf{x}_i, \text{proj}_X(\mathbf{v}) \rangle = \langle \mathbf{x}_i, \mathbf{v} \rangle$, so

$$\text{proj}_X(\text{proj}_X(\mathbf{v})) = \sum_{i=1}^m \langle \mathbf{x}_i, \text{proj}_X(\mathbf{v}) \rangle \mathbf{x}_i = \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i = \text{proj}_X(\mathbf{v}).$$

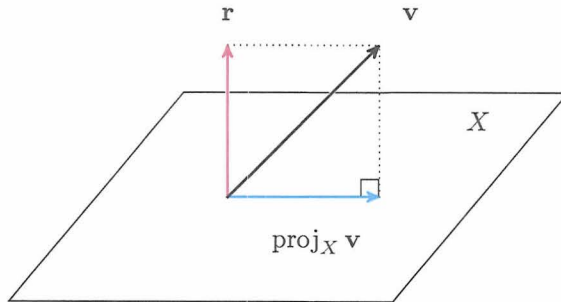


Figure 3.3. The orthogonal projection of $\mathbf{v}$ (black) onto $X = \text{span}(\{\mathbf{x}_i\}_{i=1}^m)$. The projection $\text{proj}_X \mathbf{v} = \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i$ (blue) lies in X and the difference $\mathbf{r} = \mathbf{v} - \text{proj}_X \mathbf{v}$ (red) is orthogonal to X .

(ii) Given any vector $\mathbf{x} = \sum_{i=1}^m c_i \mathbf{x}_i \in X$, we have

$$\begin{aligned} \langle \mathbf{x}, \mathbf{r} \rangle &= \left\langle \mathbf{x}, \mathbf{v} - \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i \right\rangle = \langle \mathbf{x}, \mathbf{v} \rangle - \sum_{i=1}^m \langle \mathbf{x}, \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i \rangle \\ &= \langle \mathbf{x}, \mathbf{v} \rangle - \sum_{i=1}^m \langle \mathbf{x}, \mathbf{x}_i \rangle \langle \mathbf{x}_i, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{v} \rangle - \sum_{i=1}^m \bar{c}_i \langle \mathbf{x}_i, \mathbf{v} \rangle \\ &= \langle \mathbf{x}, \mathbf{v} \rangle - \left\langle \sum_{i=1}^m c_i \mathbf{x}_i, \mathbf{v} \right\rangle = \langle \mathbf{x}, \mathbf{v} \rangle - \langle \mathbf{x}, \mathbf{v} \rangle = 0. \end{aligned}$$

(iii) Given any vector $\mathbf{x} \in X$, the square of the distance from $\mathbf{x}$ to $\mathbf{v}$ is

$$\|\mathbf{v} - \mathbf{x}\|^2 = \|\mathbf{r} + \text{proj}_X(\mathbf{v}) - \mathbf{x}\|^2 = \|\mathbf{r}\|^2 + \|\text{proj}_X(\mathbf{v}) - \mathbf{x}\|^2.$$

The last term $\|\text{proj}_X(\mathbf{v}) - \mathbf{x}\|^2$ is always nonnegative and is zero only when $\text{proj}_X(\mathbf{v}) = \mathbf{x}$, and thus $\|\mathbf{v} - \mathbf{x}\|$ is minimized when $\mathbf{x} = \text{proj}_X(\mathbf{v})$. $\square$

Remark 3.2.8. Theorem 3.2.7(iii) above shows that the linear map proj_X depends only on the subspace X and not on the particular choice of orthonormal basis $\{\mathbf{x}_i\}_{i=1}^m$.

Theorem 3.2.9 (Pythagorean Theorem). *If $\{\mathbf{x}_i\}_{i=1}^m \subset V$ is an orthonormal set with span X and $\mathbf{v} \in V$, then*

$$\|\mathbf{v}\|^2 = \|\text{proj}_X(\mathbf{v})\|^2 + \|\mathbf{v} - \text{proj}_X(\mathbf{v})\|^2 = \sum_{i=1}^m |\langle \mathbf{x}_i, \mathbf{v} \rangle|^2 + \left\| \mathbf{v} - \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{x}_i \right\|^2. \quad (3.12)$$

Proof. Since $\mathbf{v} = \text{proj}_X(\mathbf{v}) + (\mathbf{v} - \text{proj}_X(\mathbf{v}))$ and $\text{proj}_X(\mathbf{v}) \perp (\mathbf{v} - \text{proj}_X(\mathbf{v}))$, the result follows immediately by Theorems 3.1.20 and 3.2.7(ii). $\square$

Corollary 3.2.10 (Bessel's Inequality). *If $\mathbf{v} \in V$ and $\{\mathbf{x}_i\}_{i=1}^m \subset V$ is an orthonormal set with span X , then*

$$\|\mathbf{v}\|^2 \geq \|\text{proj}_X(\mathbf{v})\|^2 = \sum_{i=1}^m |\langle \mathbf{x}_i, \mathbf{v} \rangle|^2. \quad (3.13)$$

Equality holds only if $\mathbf{v} \in X$.

Proof. This follows immediately from (3.12). $\square$

3.2.2 Orthonormal Transformations

Definition 3.2.11. *A linear map L from an inner product space $(V, \langle \cdot, \cdot \rangle_V)$ to an inner product space $(W, \langle \cdot, \cdot \rangle_W)$ is called an orthonormal transformation if for every $\mathbf{x}, \mathbf{y} \in V$ we have*

$$\langle \mathbf{x}, \mathbf{y} \rangle_V = \langle L(\mathbf{x}), L(\mathbf{y}) \rangle_W. \quad (3.14)$$

If L is an orthonormal transformation from an inner product space $(V, \langle \cdot, \cdot \rangle)$ to itself, then L is called an orthonormal operator.

Proposition 3.2.12. *If L is an orthonormal operator and V is finite dimensional, then L is invertible.*

Proof. By Corollary 2.3.11, it suffices to show that L is injective. If $\mathbf{x} \in \mathcal{N}(L)$, then $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \langle L(\mathbf{x}), L(\mathbf{x}) \rangle = \|L(\mathbf{x})\|^2 = \|\mathbf{0}\|^2$, and hence $\mathbf{x} = \mathbf{0}$. $\square$

In the hypothesis of the previous proposition, it is important that the space V be finite dimensional. Otherwise, injectivity does not imply surjectivity. The following is an example of an orthonormal operator that is not invertible.

Example 3.2.13. Assume ℓ^2 is endowed with the standard inner product (3.6). Let $L : \ell^2 \rightarrow \ell^2$ be the right-shift operator given by $(a_1, a_2, \dots) \mapsto (0, a_1, a_2, \dots)$. It is easy to see that L is an injective orthonormal operator, but it is not surjective, and therefore not invertible.

Definition 3.2.14. A matrix $Q \in M_n(\mathbb{F})$ is called orthonormal¹⁹ if it is the matrix representation of an orthonormal operator on $\mathbb{F}^n$ in the standard bases (see Example 1.2.14(ii)) and with the standard inner product (see Example 3.1.5).

Theorem 3.2.15. The matrix $Q \in M_n(\mathbb{F})$ is orthonormal if and only if it satisfies $Q^H Q = Q Q^H = I$. Moreover, for any orthonormal matrix $Q \in M_n(\mathbb{F})$, the following hold:

- (i) $\|Q\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in V$.
- (ii) Q^{-1} exists and is an orthonormal matrix.
- (iii) The columns of Q are orthonormal.
- (iv) $|\det(Q)| = 1$; moreover, if $Q \in M_n(\mathbb{R})$, then $\det(Q) = \pm 1$.

Proof. See Exercise 3.10. $\square$

Corollary 3.2.16. If $Q_1, Q_2 \in M_n(\mathbb{F})$ are orthonormal matrices, then so is their product $Q_1 Q_2$.

Remark 3.2.17. An immediate consequence of Theorem 3.2.15 and Exercise 3.4 is that orthonormal matrices preserve both lengths and angles.

¹⁹Many textbooks use the term *unitary* for complex-valued orthonormal square matrices, and they call real-valued orthonormal matrices *orthogonal* matrices. In this text we prefer to use the term *orthonormal* for both of these because it more accurately describes the essential nature of these matrices, and it allows us to treat the real and complex cases identically.

Nota Bene 3.2.18. One way to determine if a given linear operator on $\mathbb{F}^n$ is orthonormal (with the usual inner product) is to represent it as a matrix in the standard basis. If that matrix is orthonormal, then the linear operator is orthonormal.

3.3 Gram–Schmidt Orthonormalization

In this section, we show that given a linearly independent set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ we can construct an orthonormal set $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$, which has the same span. An important corollary of this is that every finite-dimensional inner product space is isomorphic to $\mathbb{F}^n$ with the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y}$. This means that many results that hold on $\mathbb{F}^n$ with the standard inner product also hold for any n -dimensional inner product space (over the same field). We demonstrate this idea with a very brief treatment of the theory of orthogonal polynomials. We conclude this section by introducing the very important *QR decomposition*, which factors a matrix into a product of an orthogonal matrix and an upper-triangular matrix.

3.3.1 Gram–Schmidt Orthonormalization Process

Theorem 3.3.1 (Gram–Schmidt). *Let $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a linearly independent set in the inner product space $(V, \langle \cdot, \cdot \rangle)$. The following algorithm, called the Gram–Schmidt process, produces a set $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ that is orthonormal in V with the same span as $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.*

Let

$$\mathbf{q}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|},$$

and define $\mathbf{q}_2, \mathbf{q}_3, \dots, \mathbf{q}_n$, recursively, by

$$\mathbf{q}_k = \frac{\mathbf{x}_k - \mathbf{p}_{k-1}}{\|\mathbf{x}_k - \mathbf{p}_{k-1}\|}, \quad k = 2, \dots, n,$$

where

$$\mathbf{p}_{k-1} = \text{proj}_{Q_{k-1}}(\mathbf{x}_k) = \sum_{i=1}^{k-1} \langle \mathbf{q}_i, \mathbf{x}_k \rangle \mathbf{q}_i \quad (3.15)$$

is the projection of $\mathbf{x}_k$ onto $Q_{k-1} = \text{span}(\{\mathbf{q}_1, \dots, \mathbf{q}_{k-1}\})$.

Proof. We prove inductively that $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ is an orthonormal set. The base case is trivial. It suffices to show that if $\{\mathbf{q}_1, \dots, \mathbf{q}_{k-1}\}$ is an orthonormal set, then $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ is also an orthonormal set. By Theorem 3.2.7(ii), the residual vector $\mathbf{x}_k - \mathbf{p}_{k-1}$ is orthogonal to every vector in $Q_{k-1} = \text{span}(\{\mathbf{q}_1, \dots, \mathbf{q}_{k-1}\})$. If $\mathbf{x}_k - \mathbf{p}_{k-1} = \mathbf{0}$, then $\mathbf{x}_k = \mathbf{p}_{k-1} \in Q_{k-1}$, which contradicts the linear independence of $\{\mathbf{q}_1, \dots, \mathbf{q}_{k-1}\}$ given by the inductive hypothesis. Thus, $\mathbf{x}_k - \mathbf{p}_{k-1}$ is nonzero and $\mathbf{q}_k$ is well defined. It follows that $\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$ is an orthonormal set, as required. $\square$

Example 3.3.2. Consider the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

in $\mathbb{R}^3$. The Gram–Schmidt process yields

$$\mathbf{q}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

Projecting $\mathbf{x}_2$ onto $\mathbf{q}_1$ gives

$$\mathbf{p}_1 = \langle \mathbf{q}_1, \mathbf{x}_2 \rangle \mathbf{q}_1 = \frac{2}{3} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{q}_2 = \frac{\mathbf{x}_2 - \mathbf{p}_1}{\|\mathbf{x}_2 - \mathbf{p}_1\|} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

Now projecting onto $Q_2 = \text{span}\{\mathbf{q}_1, \mathbf{q}_2\}$ gives

$$\mathbf{p}_2 = \langle \mathbf{q}_1, \mathbf{x}_3 \rangle \mathbf{q}_1 + \langle \mathbf{q}_2, \mathbf{x}_3 \rangle \mathbf{q}_2 = \frac{2}{3} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + \frac{5}{6} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}$$

and

$$\mathbf{q}_3 = \frac{\mathbf{x}_3 - \mathbf{p}_2}{\|\mathbf{x}_3 - \mathbf{p}_2\|} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Thus, we have the following orthonormal basis for $\mathbb{R}^3$:

$$\left\{ \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Example 3.3.3. Consider the inner product space $\mathbb{F}[x; 2]$ with the inner product

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx. \quad (3.16)$$

We apply the Gram–Schmidt process to the set of vectors $\{1, x, x^2\} \subset \mathbb{F}[x; 2]$. The first step yields

$$q_1 = \frac{1}{\|1\|} = \frac{1}{\sqrt{2}} \quad \text{and} \quad p_1 = \left\langle \frac{1}{\sqrt{2}}, x \right\rangle \frac{1}{\sqrt{2}} = 0.$$

Thus, x is already orthogonal to q_1 . The second step gives

$$q_2 = \sqrt{\frac{3}{2}}x \quad \text{and} \quad p_2 = \left\langle \frac{1}{\sqrt{2}}, x^2 \right\rangle \frac{1}{\sqrt{2}} + \left\langle \sqrt{\frac{3}{2}}x, x^2 \right\rangle \sqrt{\frac{3}{2}}x = \frac{1}{3}.$$

It follows that

$$q_3 = \frac{x^2 - \frac{1}{3}}{\|x^2 - \frac{1}{3}\|} = \sqrt{\frac{5}{8}}(3x^2 - 1).$$

Vista 3.3.4. In Example 3.3.3, we used the Gram–Schmidt method to derive the orthonormal polynomials $\{q_1, q_2, q_3\}$ spanning $\mathbb{F}[x; 2]$. Repeating this for all degrees gives the *Legendre polynomials*.

There are many other inner products that could be used on $\mathbb{F}[x]$. For example, consider the weighted inner product

$$\langle f, g \rangle = \int_a^b w(x)f(x)g(x)dx, \quad (3.17)$$

where $w(x) > 0$ is a weight function. By performing the Gram–Schmidt orthogonalization process, we can construct a collection of polynomials that are orthonormal with respect to (3.17). In the table below, we list a few of the domains, weights, and the names of the corresponding families of orthogonal polynomials. These are discussed in much more detail in Volume 2.

Name	Domain	$w(x)$
Chebyshev-1	$(-1, 1)$	$(1 - x^2)^{-1/2}$
Chebyshev-2	$(-1, 1)$	$(1 - x^2)^{1/2}$
Hermite	$(-\infty, \infty)$	$\exp(-x^2)$
Legendre	$(-1, 1)$	1

Remark 3.3.5. Naïve application of the Gram–Schmidt process, as described above, is numerically unstable, meaning that the round-off error in numerical computation compounds to yield unreliable output. A slight change gives the *modified Gram–Schmidt* routine and produces better results; see the lab on QR decomposition for details.

3.3.2 Finite-Dimensional Inner Product Spaces

Corollary 3.3.6. If $(V, \langle \cdot, \cdot \rangle_V)$ is an n -dimensional inner product space over $\mathbb{F}$, then it is isomorphic to $\mathbb{F}^n$ with the standard inner product (3.2).

Proof. Using Theorem 3.3.1, we choose an orthonormal basis $[\mathbf{x}_1, \dots, \mathbf{x}_n]$ for V and define the map $L : V \rightarrow \mathbb{F}^n$ by

$$L\left(\sum_{i=1}^n a_i \mathbf{x}_i\right) = [a_1 \quad \dots \quad a_n]^\top \in \mathbb{F}^n.$$

This is clearly a bijective linear transformation. To see that the inner product is preserved, we note that if $\mathbf{x} = \sum_{j=1}^n b_j \mathbf{x}_j$ and $\mathbf{y} = \sum_{k=1}^n c_k \mathbf{x}_k$, then

$$\langle \mathbf{x}, \mathbf{y} \rangle_V = \left\langle \sum_{j=1}^n b_j \mathbf{x}_j, \sum_{k=1}^n c_k \mathbf{x}_k \right\rangle_V = \sum_{j=1}^n \sum_{k=1}^n \overline{b_j} c_k \langle \mathbf{x}_j, \mathbf{x}_k \rangle_V = \sum_{j=1}^n \overline{b_j} c_j = \langle L(\mathbf{x}), L(\mathbf{y}) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{F}^n$. $\square$

Remark 3.3.7. Just as Corollary 2.3.12 is a big deal for vector spaces, it follows that Corollary 3.3.6 is a big deal for inner product spaces. It means that although there are many different *descriptions* of finite-dimensional inner product spaces, there is essentially (up to isomorphism) *only one n -dimensional inner product space* over $\mathbb{F}$ for each $n \in \mathbb{N}$. Anything we can prove on $\mathbb{F}^n$ with the standard inner product automatically holds for any finite-dimensional inner product space over $\mathbb{F}$.

Example 3.3.8. Recall that the orthonormal basis $\{q_1, q_2, q_3\}$, computed in Example 3.3.3, has the following transition matrix:

$$[1, x, x^2] = [q_1, q_2, q_3] \frac{\sqrt{2}}{3} \begin{bmatrix} 3 & 0 & 1 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \frac{2}{\sqrt{5}} \end{bmatrix}.$$

Consider the linear isomorphism $L : \mathbb{F}[x; 2] \rightarrow \mathbb{F}^3$, given by $L(q_i) = \mathbf{e}_i$. Since $[q_1, q_2, q_3]$ is orthonormal, this map also preserves the inner product, so $\langle f, g \rangle_{\mathbb{F}[x; 2]} = \langle L(f), L(g) \rangle_{\mathbb{F}^3}$.

To express this map in terms of the power basis $[1, x, x^2]$, first write an arbitrary element of $\mathbb{F}[x; 2]$ as $p(x) = ax^2 + bx + c = [1, x, x^2] \begin{bmatrix} c & b & a \end{bmatrix}^\top$, and then compute $L(p)$ as

$$\begin{aligned} L(p) &= [L(1), L(x), L(x^2)] \begin{bmatrix} c \\ b \\ a \end{bmatrix} = [L(q_1), L(q_2), L(q_3)] \frac{\sqrt{2}}{3} \begin{bmatrix} 3 & 0 & 1 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} \\ &= [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3] \frac{\sqrt{2}}{3} \begin{bmatrix} 3c + a \\ \sqrt{3}b \\ \frac{2}{\sqrt{5}}a \end{bmatrix}. \end{aligned}$$

3.3.3 The QR Decomposition

The QR decomposition of a matrix is an important tool that allows any matrix to be written as a product of an orthonormal matrix Q and an upper-triangular

matrix R . The QR decomposition is used in many applications, such as solving least squares problems and computing eigenvalues of a matrix, and it is one of the most fundamental matrix decompositions in applied and computational mathematics. Here we prove the existence of the decomposition using the Gram–Schmidt orthonormalization process.

Theorem 3.3.9 (QR Decomposition). *Any matrix $A \in M_{m \times n}$ of rank $n \leq m$ can be factored into a product $A = QR$, where Q is an $m \times m$ orthonormal matrix and R is an $m \times n$ upper-triangular matrix.*

Proof. Since A has rank n , the columns of A are linearly independent, so we can apply Theorem 3.3.1. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the columns of A . By the replacement theorem (Theorem 1.4.1) there are vectors $\mathbf{x}_{n+1}, \dots, \mathbf{x}_m$ such that space $\mathbb{F}^m$ is spanned by $\mathbf{x}_1, \dots, \mathbf{x}_m$.

Using the notation in Theorem 3.3.1, let $\mathbf{p}_0 = \mathbf{0}$, let $r_{jk} = \langle \mathbf{q}_j, \mathbf{x}_k \rangle$ for $j = 1, 2, \dots, k-1$, and let $r_{kk} = \|\mathbf{x}_k - \mathbf{p}_{k-1}\|$. We have

$$\begin{aligned} \mathbf{x}_1 &= r_{11}\mathbf{q}_1, \\ \mathbf{x}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2, \\ &\vdots \\ \mathbf{x}_m &= r_{1m}\mathbf{q}_1 + r_{2m}\mathbf{q}_2 + \cdots + r_{mm}\mathbf{q}_m, \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_m \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ 0 & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{mm} \end{bmatrix}.$$

Let Q be the orthonormal matrix $[\mathbf{q}_1, \dots, \mathbf{q}_m]$, and let R be the first n columns of the upper-triangular matrix $[r_{ij}]$ above. Since the columns of A are $\mathbf{x}_1, \dots, \mathbf{x}_n$, this gives $A = QR$, as required. $\square$

Remark 3.3.10. Along with the full QR decomposition given in Theorem 3.3.9, another useful decomposition is the *reduced QR decomposition*. Given a matrix $A \in M_{m \times n}$ of rank $n \leq m$, the reduced QR decomposition of A is the factorization $A = \hat{Q}\hat{R}$, where $\hat{Q}$ is an $m \times n$ orthonormal matrix and $\hat{R}$ is an $n \times n$ upper-triangular matrix.

Example 3.3.11. To compute the reduced QR decomposition of

$$A = \begin{bmatrix} 1 & -2 & 3.5 \\ 1 & 3 & -0.5 \\ 1 & 3 & 2.5 \\ 1 & -2 & 0.5 \end{bmatrix}$$

let $\mathbf{x}_i$ denote the i th column of the matrix A , so that

$$\mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T, \quad \mathbf{x}_2 = \begin{bmatrix} -2 & 3 & 3 & -2 \end{bmatrix}^T, \quad \text{and} \quad \mathbf{x}_3 = \frac{1}{2} \begin{bmatrix} 7 & -1 & 5 & 1 \end{bmatrix}^T.$$

Begin by normalizing $\mathbf{x}_1$. This gives $r_{11} = \|\mathbf{x}_1\| = 2$ and

$$\mathbf{q}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T.$$

Next find $r_{12} = \langle \mathbf{q}_1, \mathbf{x}_2 \rangle = 1$ and $r_{13} = \langle \mathbf{q}_1, \mathbf{x}_3 \rangle = 3$. Now compute

$$\mathbf{x}_2 - \mathbf{p}_1 = \mathbf{x}_2 - r_{12}\mathbf{q}_1 = \frac{5}{2} \begin{bmatrix} -1 & 1 & 1 & -1 \end{bmatrix}^T.$$

Normalizing gives $r_{22} = \|\mathbf{x}_2 - \mathbf{p}_1\| = 5$ and $\mathbf{q}_2 = \frac{1}{5} \begin{bmatrix} -1 & 1 & 1 & -1 \end{bmatrix}^T$. Further, we find $r_{23} = \langle \mathbf{q}_2, \mathbf{x}_3 \rangle = -1$. Finally, we have

$$\mathbf{x}_3 - \mathbf{p}_2 = \mathbf{x}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2 = \frac{3}{2} \begin{bmatrix} 1 & -1 & 1 & -1 \end{bmatrix}^T.$$

This gives $r_{33} = 3$ and $\mathbf{q}_3 = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 & -1 \end{bmatrix}^T$. Therefore, $\widehat{Q}$ is the matrix

$$\widehat{Q} = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{bmatrix}.$$

You should verify that $\widehat{Q}$ has orthonormal columns, that is, $\widehat{Q}^T \widehat{Q} = I$. We also have

$$\widehat{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} \|\mathbf{x}_1\| & \langle \mathbf{q}_1, \mathbf{x}_2 \rangle & \langle \mathbf{q}_1, \mathbf{x}_3 \rangle \\ 0 & \|\mathbf{x}_2 - \mathbf{p}_1\| & \langle \mathbf{q}_2, \mathbf{x}_3 \rangle \\ 0 & 0 & \|\mathbf{x}_3 - \mathbf{p}_2\| \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & -1 \\ 0 & 0 & 3 \end{bmatrix}.$$

For the full QR decomposition, we can choose any additional $\mathbf{x}_4$ such that $[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4]$ spans $\mathbb{R}^4$, and then continue the Gram-Schmidt process. A convenient choice in this example is $\mathbf{x}_4 = \mathbf{e}_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$. We calculate $r_{14} = \langle \mathbf{q}_1, \mathbf{x}_4 \rangle = 1/2$ and $r_{24} = r_{34} = -1/2$. This gives $\mathbf{p}_4 = \frac{1}{2}(\mathbf{q}_1 - \mathbf{q}_2 - \mathbf{q}_3)$ and $\mathbf{x}_4 - \mathbf{p}_4 = \frac{1}{4} \begin{bmatrix} -1 & -1 & 1 & 1 \end{bmatrix}^T$, from which we find $\mathbf{q}_4 = \frac{1}{2} \begin{bmatrix} -1 & -1 & 1 & 1 \end{bmatrix}^T$. Thus for the full QR decomposition, we have

$$Q = \frac{1}{2} \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

and

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & -1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Vista 3.3.12. The QR decomposition is an important tool for solving many fundamental problems. Two of the most important applications of the QR decomposition are in solving least squares problems (see Section 3.9 and Vista 3.9.8) and in finding eigenvalues (see Section 13.4.2).

Remark 3.3.13. The QR decomposition can also be used to solve a linear system of the form $A\mathbf{x} = \mathbf{b}$. By writing $A = QR$ we have $QR\mathbf{x} = \mathbf{b}$, and since Q is orthonormal, we have

$$R\mathbf{x} = Q^T QR\mathbf{x} = Q^T \mathbf{b}.$$

Since R is upper triangular, the system $R\mathbf{x} = Q^T \mathbf{b}$ can be backsolved to find $\mathbf{x}$. This method takes about twice as many operations as Gaussian elimination, but is more stable. In practice, however, the cases where Gaussian elimination is unstable are extremely rare, and so Gaussian elimination is usually preferred for solving dense linear systems.

3.4 QR with Householder Transformations

There are a few different algorithms for computing the QR decomposition. We have already seen how the Gram–Schmidt (or modified Gram–Schmidt) process can be used to compute the QR decomposition. In this section we present an algorithm using what are called *Householder transformations*. This algorithm is usually faster and more accurate than the algorithms using Gram–Schmidt. Householder transformations modify $A \in M_{m \times n}(\mathbb{F})$ to produce an upper-triangular matrix $R \in M_{m \times n}(\mathbb{F})$ through a sequence of left multiplications of orthonormal matrices $Q_i^H \in M_n(\mathbb{F})$. These matrices are then multiplied together to form $Q^H \in M_n(\mathbb{F})$. In other words, $R = Q_k^H \cdots Q_2^H Q_1^H A = Q^H A$, which implies that $A = QR$.

3.4.1 The Geometry of Householder Transformations

In order to construct Householder transformations, we first need to know how to project vectors onto the orthogonal complement of a given nonzero vector.

Definition 3.4.1. Given a nonzero vector $\mathbf{v} \in \mathbb{F}^n$, the orthogonal complement of $\mathbf{v}$, denoted $\mathbf{v}^\perp$, is the set of all vectors $\mathbf{x} \in \mathbb{F}^n$ that are orthogonal to $\mathbf{v}$. In other words, $\mathbf{v}^\perp = \{\mathbf{x} \in \mathbb{F}^n \mid \langle \mathbf{v}, \mathbf{x} \rangle = 0\}$.

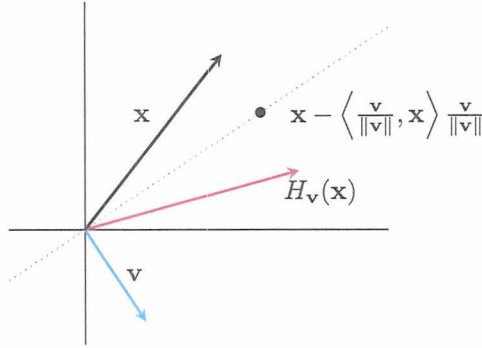


Figure 3.4. Reflection $H_v(\mathbf{x})$ (red) of $\mathbf{x}$ (black) through the orthogonal complement (dotted) of $\mathbf{v}$ (blue).

Nota Bene 3.4.2. Beware that although $\mathbf{v}$ is a vector, $\mathbf{v}^\perp$ is a set (in fact a whole subspace)—not a vector.

Recall that $\text{proj}_v(\mathbf{x})$ is the projection of $\mathbf{x}$ onto the nonzero vector $\mathbf{v}$, and thus the residual $\mathbf{x} - \text{proj}_v(\mathbf{x})$ is orthogonal to $\mathbf{v}$; see Proposition 3.1.24(ii) for details. This means that we can project an arbitrary vector $\mathbf{x} \in \mathbb{F}^n$ onto $\mathbf{v}^\perp$ by the map

$$\text{proj}_{\mathbf{v}^\perp}(\mathbf{x}) = \mathbf{x} - \text{proj}_v(\mathbf{x}) = \mathbf{x} - \frac{\mathbf{v}}{\|\mathbf{v}\|^2} \langle \mathbf{v}, \mathbf{x} \rangle = \left(I - \frac{\mathbf{v}\mathbf{v}^H}{\mathbf{v}^H\mathbf{v}} \right) \mathbf{x}.$$

A Householder transformation (also called a Householder reflection) reflects the vector $\mathbf{x}$ across the orthogonal complement $\mathbf{v}^\perp$, in essence moving twice as far as the projection; see Figure 3.4 for an illustration of the Householder reflection.

Definition 3.4.3. Fix a nonzero vector $\mathbf{v} \in \mathbb{F}^n$. For any $\mathbf{x} \in \mathbb{F}^n$ the Householder transformation reflecting $\mathbf{x}$ across the orthogonal complement $\mathbf{v}^\perp$ of $\mathbf{v}$ is given by

$$H_v(\mathbf{x}) = \mathbf{x} - 2 \text{proj}_v(\mathbf{x}) = \left(I - 2 \frac{\mathbf{v}\mathbf{v}^H}{\mathbf{v}^H\mathbf{v}} \right) \mathbf{x}. \quad (3.18)$$

Proposition 3.4.4. Assume $\mathbf{v}$ in $\mathbb{F}^n$ is nonzero. We have the following properties of Householder reflections:

- (i) H_v is an orthonormal transformation.
- (ii) Elements in $\mathbf{v}^\perp$ are not affected by H_v . More precisely, if $\langle \mathbf{v}, \mathbf{x} \rangle = 0$, then $H_v(\mathbf{x}) = \mathbf{x}$.

Proof. The proof is Exercise 3.20. $\square$

3.4.2 Computing the QR Decomposition via Householder

We now compute the QR decomposition using Householder reflections. Given an arbitrary vector $\mathbf{x}$ in $\mathbb{F}^n$, we want to find the nonzero vector $\mathbf{v} \in \mathbb{F}^n$ whose Householder reflection $H_{\mathbf{v}}(\mathbf{x})$ resides in the span of the first standard basis vector $\mathbf{e}_1$. The following lemmata²⁰ tell us how to find $\mathbf{v}$. We begin with real-valued matrices and then generalize to complex-valued matrices.

Lemma 3.4.5. *Given $\mathbf{x} \in \mathbb{R}^n$, if the vector $\mathbf{v} = \mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1$ is nonzero, then $H_{\mathbf{v}}(\mathbf{x}) = -\|\mathbf{x}\|\mathbf{e}_1$. Similarly, if $\mathbf{v} = \mathbf{x} - \|\mathbf{x}\|\mathbf{e}_1$ is nonzero, then $H_{\mathbf{v}}(\mathbf{x}) = \|\mathbf{x}\|\mathbf{e}_1$.*

Proof. We prove the case that $\mathbf{v} = \mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1$ is nonzero. The proof for $\mathbf{v} = \mathbf{x} - \|\mathbf{x}\|\mathbf{e}_1$ is similar. Let x_1 denote the first component of $\mathbf{x}$. Expanding the definition of $H_{\mathbf{v}}(\mathbf{x})$, we have

$$\begin{aligned} H_{\mathbf{v}}(\mathbf{x}) &= \mathbf{x} - 2 \frac{(\mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1)(\mathbf{x}^T + \|\mathbf{x}\|\mathbf{e}_1^T)}{(\mathbf{x}^T + \|\mathbf{x}\|\mathbf{e}_1^T)(\mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1)} \mathbf{x} \\ &= \mathbf{x} - \frac{\|\mathbf{x}\|^2 \mathbf{x} + \|\mathbf{x}\|^3 \mathbf{e}_1 + \|\mathbf{x}\|x_1 \mathbf{x} + \|\mathbf{x}\|^2 x_1 \mathbf{e}_1}{\|\mathbf{x}\|^2 + x_1 \|\mathbf{x}\|} \\ &= \mathbf{x} - \frac{(\|\mathbf{x}\| + x_1)\mathbf{x} + \|\mathbf{x}\|(x_1 + \|\mathbf{x}\|)\mathbf{e}_1}{\|\mathbf{x}\| + x_1} \\ &= -\|\mathbf{x}\|\mathbf{e}_1. \quad \square \end{aligned}$$

Remark 3.4.6. Mathematically, it is enough to require $\mathbf{v}$ to be nonzero when doing Householder transformations. But computationally, with floating-point arithmetic, we don't want $\mathbf{v}$ to be at all close to zero, since the operator $H_{\mathbf{v}}$ becomes much more sensitive to round-off error as $\|\mathbf{v}\|$ gets smaller. To reduce the numerical errors, when $\mathbf{x} \approx \|\mathbf{x}\|\mathbf{e}_1$, we choose the “plus” option in Lemma 3.4.5 and set $\mathbf{v} = \mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1$. Similarly, when $\mathbf{x} \approx -\|\mathbf{x}\|\mathbf{e}_1$, we choose the “minus” option and set $\mathbf{v} = \mathbf{x} - \|\mathbf{x}\|\mathbf{e}_1$. We can succinctly combine these choices into the single rule $\mathbf{v} = \mathbf{x} + \text{sign}(x_1)\|\mathbf{x}\|\mathbf{e}_1$, which automatically chooses the option farther from zero. Here we use the convention that $\text{sign}(0) = 1$.

Lemma 3.4.7. *Given $\mathbf{x} \in \mathbb{C}^n$, let x_1 denote the first component of $\mathbf{x}$. Recall²¹ that the complex sign of x_1 is $\text{sign}(x_1) = x_1/|x_1|$ when $x_1 \neq 0$, and $\text{sign}(0) = 1$. Choosing $\mathbf{v} = \mathbf{x} + \text{sign}(x_1)\|\mathbf{x}\|\mathbf{e}_1$ implies that*

$$H_{\mathbf{v}}(\mathbf{x}) = -\text{sign}(x_1)\|\mathbf{x}\|\mathbf{e}_1$$

has zeros in all but the first entry.

Proof. The proof is Exercise 3.21. $\square$

²⁰Lemmata is the plural of the Greek word *lemma*.

²¹See (B.3) in Appendix B.1.2.

Remark 3.4.8. The vector $\mathbf{v} = \mathbf{x} + \text{sign}(x_1)\|\mathbf{x}\|\mathbf{e}_1$ is never zero unless $\mathbf{x} = \mathbf{0}$. Indeed if there exists a nonzero $\mathbf{x}$ such that $\mathbf{x} = -\text{sign}(x_1)\|\mathbf{x}\|\mathbf{e}_1$, then $x_1 = -\text{sign}(x_1)\|\mathbf{x}\|$. This implies that $\text{sign}(x_1) = -\text{sign}(x_1)$, which is a contradiction, since $\text{sign}(x_1)$ is never zero.

The algorithm for computing the QR decomposition with Householder reflections is essentially just a repeated application of Lemma 3.4.7, as we show in the next theorem.

Theorem 3.4.9. *Given a matrix $A \in M_{m \times n}(\mathbb{F})$, with $m \geq n$, there is a sequence of nonzero vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_\ell \in \mathbb{F}^m$, with $\ell = \min\{m-1, n\}$, such that $R = H_{\mathbf{v}_\ell} H_{\mathbf{v}_{\ell-1}} \cdots H_{\mathbf{v}_1} A$ is upper triangular. Taking $Q = H_{\mathbf{v}_1}^\mathbf{H} H_{\mathbf{v}_2}^\mathbf{H} \cdots H_{\mathbf{v}_\ell}^\mathbf{H}$ gives a QR decomposition of A .*

Proof. Let $\mathbf{x}_1$ be the first column of the matrix A . Following Lemma 3.4.7, set $\mathbf{v}_1 = \mathbf{x}_1 + e^{i\theta_1}\|\mathbf{x}_1\|\mathbf{e}_1$, where $e^{i\theta_1} = \text{sign}(x_1)$, so that the Householder reflection $H_{\mathbf{v}_1}$ takes $\mathbf{x}_1$ to the span of $\mathbf{e}_1$. Therefore, $H_{\mathbf{v}_1}A$ has the form

$$H_{\mathbf{v}_1}A = \begin{bmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & * \cdots & * \end{bmatrix},$$

where $*$ indicates an entry of arbitrary value.

Now let $\mathbf{x}_2$ be the second column of the new matrix $H_{\mathbf{v}_1}A$. We decompose $\mathbf{x}_2$ into $\mathbf{x}_2 = \mathbf{x}'_2 + \mathbf{x}''_2$, where $\mathbf{x}'_2$ and $\mathbf{x}''_2$ are of the form

$$\mathbf{x}'_2 = [z_1 \ 0 \ \cdots \ 0]^\mathbf{T} \quad \text{and} \quad \mathbf{x}''_2 = [0 \ z_2 \ \cdots \ z_m]^\mathbf{T}.$$

Set $\mathbf{v}_2 = \mathbf{x}''_2 + e^{i\theta_2}\|\mathbf{x}''_2\|\mathbf{e}_2$. Since $\mathbf{x}''_2$ and $\mathbf{e}_2$ are both orthogonal to $\mathbf{e}_1$, the vector $\mathbf{v}_2$ is also orthogonal to $\mathbf{e}_1$. Thus, the reflection $H_{\mathbf{v}_2}$ acts as the identity on the span of $\mathbf{e}_1$, and $H_{\mathbf{v}_2}$ leaves $\mathbf{x}'_2$ and the first column of $H_{\mathbf{v}_1}A$ unchanged.

We now have $H_{\mathbf{v}_2}\mathbf{x}_2 = H_{\mathbf{v}_2}\mathbf{x}'_2 + H_{\mathbf{v}_2}\mathbf{x}''_2 = \mathbf{x}'_2 + H_{\mathbf{v}_2}\mathbf{x}''_2$. By Lemma 3.4.7 the vector $H_{\mathbf{v}_2}\mathbf{x}''_2$ lies in $\text{span}\{\mathbf{e}_2\}$; so $H_{\mathbf{v}_2}\mathbf{x}_2 \in \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$, and the matrix $H_{\mathbf{v}_2}H_{\mathbf{v}_1}A$ has the form

$$H_{\mathbf{v}_2}H_{\mathbf{v}_1}A = \begin{bmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & * & * \cdots & * \end{bmatrix}.$$

Repeat the procedure for each $k = 3, 4, \dots, \ell$, choosing $\mathbf{x}_k$ equal to the k th column of $H_{\mathbf{v}_{k-1}} \cdots H_{\mathbf{v}_2}H_{\mathbf{v}_1}A$, and decomposing $\mathbf{x}_k = \mathbf{x}'_k + \mathbf{x}''_k$, where $\mathbf{x}'_k$ and $\mathbf{x}''_k$ are of the form

$$\mathbf{x}'_k = [z_1 \ \cdots \ z_{k-1} \ 0 \ \cdots \ 0]^\mathbf{T} \quad \text{and} \quad \mathbf{x}''_k = [0 \ \cdots \ 0 \ z_k \ \cdots \ z_m]^\mathbf{T}.$$

Set $\mathbf{v}_k = \mathbf{x}_k'' + e^{i\theta_k} \|\mathbf{x}_k''\| \mathbf{e}_k$. The vectors $\mathbf{x}_k''$ and $\mathbf{e}_k$ are both orthogonal to the subspace spanned by $\mathbf{e}_1, \dots, \mathbf{e}_{k-1}$, so $\mathbf{v}_k$ is as well. This implies that the reflection $H_{\mathbf{v}_k}$ acts as the identity on the span of $\mathbf{e}_1, \dots, \mathbf{e}_{k-1}$. In other words, since the first $k-1$ columns of $H_{\mathbf{v}_{k-1}} \cdots H_{\mathbf{v}_2} H_{\mathbf{v}_1} A$ are in upper-triangular form, they remain unchanged by $H_{\mathbf{v}_k}$. Moreover, we have $H_{\mathbf{v}_k} \mathbf{x}_k = \mathbf{x}_k' + H_{\mathbf{v}_k} \mathbf{x}_k''$. By Lemma 3.4.7 we have that $H_{\mathbf{v}_k} \mathbf{x}_k'' \in \text{span}\{\mathbf{e}_k\}$, and so $H_{\mathbf{v}_k} \mathbf{x}_k \in \text{span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k\}$, as desired.

Upon termination, we set $R = H_{\mathbf{v}_\ell} \cdots H_{\mathbf{v}_2} H_{\mathbf{v}_1} A$, noting that it is upper triangular. The QR decomposition follows by taking $Q = H_{\mathbf{v}_1}^H H_{\mathbf{v}_2}^H \cdots H_{\mathbf{v}_\ell}^H$. By Proposition 3.4.4 each $H_{\mathbf{v}_k}$ is orthonormal, so the product Q is as well. $\square$

Remark 3.4.10. Recall that at the k th reflection, the first $k-1$ columns are unaffected by $H_{\mathbf{v}_k}$. It turns out that the first $k-1$ rows are also unaffected by $H_{\mathbf{v}_k}$. This is harder to see than for columns, but it follows from the fact that the k th Householder transform is block diagonal with the $(1,1)$ block being the $(k-1) \times (k-1)$ identity. The numerical algorithms for computing the QR decomposition take advantage of these facts and skip those parts of the calculations. This speeds up these algorithms considerably.

3.4.3 A Complete Worked Example

Consider the matrix from Example 3.3.11:

$$A = \begin{bmatrix} 1 & -2 & 3.5 \\ 1 & 3 & -0.5 \\ 1 & 3 & 2.5 \\ 1 & -2 & 0.5 \end{bmatrix}.$$

To compute the QR decomposition using Householder reflections, let $\mathbf{x}_1$ be the first column of A and set

$$\mathbf{v}_1 = \mathbf{x}_1 - \text{sign}(x_{11}) \|\mathbf{x}_1\| \mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - 2 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

This gives

$$H_{\mathbf{v}_1} = I - 2 \frac{\mathbf{v}_1 \mathbf{v}_1^H}{\mathbf{v}_1^H \mathbf{v}_1} = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 & -1/2 \\ 1/2 & -1/2 & -1/2 & 1/2 \end{bmatrix},$$

which yields

$$H_{\mathbf{v}_1} A = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & -5 & 1 \end{bmatrix}.$$

Let $\mathbf{x}_2$ be the second column of $H_{\mathbf{v}_1} A$, noting that

$$\mathbf{x}_2' = [1 \ 0 \ 0 \ 0]^T \quad \text{and} \quad \mathbf{x}_2'' = [0 \ 0 \ 0 \ -5]^T.$$

Now set

$$\mathbf{v}_2 = \mathbf{x}_2'' - \text{sign}(\mathbf{x}_2'')\|\mathbf{x}_2''\|\mathbf{e}_2 = \begin{bmatrix} 0 \\ -5 \\ 0 \\ -5 \end{bmatrix}.$$

This gives

$$H_{\mathbf{v}_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix},$$

which yields

$$H_{\mathbf{v}_2}H_{\mathbf{v}_1}A = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & -1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore,

$$R = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 5 & -1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$Q = (H_{\mathbf{v}_2}H_{\mathbf{v}_1})^{-1} = H_{\mathbf{v}_1}H_{\mathbf{v}_2} = \begin{bmatrix} 1/2 & -1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 & 1/2 \end{bmatrix},$$

where the second equality comes from the fact that $H_{\mathbf{v}_1}$ and $H_{\mathbf{v}_2}$ are orthonormal matrices.

3.5 Normed Linear Spaces

Up to this point we've used the notation $\|\cdot\|$ to denote the length of a vector induced by an inner product, and we did this without any justification or explanation. In this section, we show that this notion of length really does have the properties that one would expect of a length function. But there are also other ways of defining the length, or *norm*, of a vector.

Norms are useful for many things, in particular, for measuring when two vectors are close together (or far apart). This allows us to quantify the convergence of sequences, to approximate vectors with other vectors, and to give geometric meaning to properties and relations in vector spaces. Different norms give different definitions of which vectors are “close” to one another, and it is very handy to have many alternatives to choose from, depending on the problem we want to solve.

In this section, we describe the basic properties of norms and give examples of various kinds of norms, some of which can be induced by an inner product and some of which cannot.

Definition 3.5.1. A norm on a vector space V is a map $\|\cdot\| : V \rightarrow [0, \infty)$ satisfying the following conditions for all $\mathbf{x}, \mathbf{y} \in V$ and all $a \in \mathbb{F}$:

- (i) *Positivity*: $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = 0$.
- (ii) *Scale preservation*: $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$.
- (iii) *Triangle inequality*: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

If $\|\cdot\|$ satisfies all the conditions above except that $\|\mathbf{x}\| = 0$ need not imply that $\mathbf{x} = 0$, then $\|\cdot\|$ is called a *seminorm*. A vector space with a norm is called a *normed linear space* and is often denoted by the pair $(V, \|\cdot\|)$.

Theorem 3.5.2. Every inner product space is a normed linear space with norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Proof. The first two conditions follow immediately from the definition of an inner product; see Remark 3.1.12. It remains to verify the triangle inequality:

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \|\mathbf{x}\|^2 + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \|\mathbf{y}\|^2 \\ &\leq \|\mathbf{x}\|^2 + 2|\langle \mathbf{x}, \mathbf{y} \rangle| + \|\mathbf{y}\|^2 \\ &\leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \quad (\text{by Cauchy-Schwarz}) \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \end{aligned}$$

Hence, $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. $\square$

Remark 3.5.3. In most situations, the most natural norm to use on an inner product space is the induced norm $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. Whenever we are working in an inner product space, we will almost always use the induced norm, unless we explicitly say otherwise.

3.5.1 Examples

Example 3.5.4. For any $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T \in \mathbb{F}^n$, the following are norms on $\mathbb{F}^n$:

- (i) 1-norm.

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \dots + |x_n|. \quad (3.19)$$

The 1-norm is sometimes called the *taxicab norm* or the *Manhattan norm* because it tracks the distance traversed along streets in a rectilinear grid.

- (ii) 2-norm.

$$\|\mathbf{x}\|_2 = (|x_1|^2 + |x_2|^2 + \dots + |x_n|^2)^{1/2}. \quad (3.20)$$

This is the usual *Euclidean norm*, and it measures the distance that a crow would fly to get from one point to another.

- (iii) ∞ -norm.

$$\|\mathbf{x}\|_\infty = \sup\{|x_1|, |x_2|, \dots, |x_n|\}. \quad (3.21)$$

Figure 3.5 shows the sets $\{\mathbf{x} \in \mathbb{R}^2: \|\mathbf{x}\| \leq 1\}$ for the 1-norm, the 2-norm, and the ∞ -norm, and Figure 3.6 shows the analogous sets in $\mathbb{R}^3$. As mentioned above, the 2-norm arises from the standard inner product on $\mathbb{F}^n$, but neither the 1-norm nor the ∞ -norm arises from any inner product.

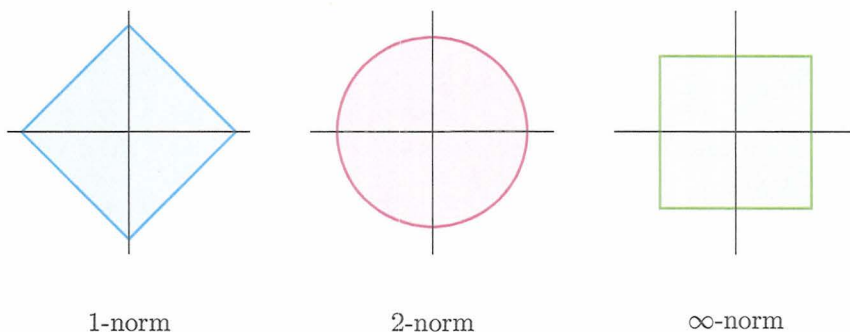


Figure 3.5. The closed unit ball $\{\mathbf{x} \in \mathbb{R}^2: \|\mathbf{x}\| \leq 1\}$ for the 1-norm, the 2-norm, and the ∞ -norm in $\mathbb{R}^2$, as discussed in Example 3.5.4.

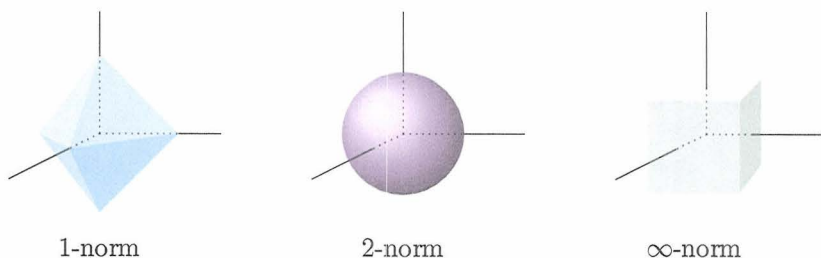


Figure 3.6. The closed unit ball $\{\mathbf{x} \in \mathbb{R}^3: \|\mathbf{x}\| \leq 1\}$ for the 1-norm, the 2-norm, and the ∞ -norm in $\mathbb{R}^3$, as discussed in Example 3.5.4.

Example 3.5.5. The 1- and 2-norms are special cases of p -norms ($p \geq 1$) given by

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}. \quad (3.22)$$

We show in Corollary 3.6.7 that the p -norm satisfies the triangle inequality.

It is not hard to show that the ∞ -norm is the limit of p -norms as $p \rightarrow \infty$; that is,

$$\|\mathbf{x}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{x}\|_p.$$

It is common to say that the ∞ -norm is the p -norm for $p = \infty$.

Example 3.5.6. The *Frobenius norm* on the space $M_{m \times n}(\mathbb{F})$ is given by

$$\|A\|_F = \sqrt{\operatorname{tr}(A^H A)}.$$

This arises from the Frobenius (or Hilbert–Schmidt) inner product (3.5). It is worth noting that the square of the Frobenius norm is just the sum of squares of the matrix elements. Thus, if you stack the elements of A into a long vector of dimension $mn \times 1$, the Frobenius norm of A is just the usual 2-norm of that stacked vector.

The Frobenius norm holds a prominent place in applied linear algebra, in part because it is easy to compute, but also because it is invariant under orthonormal transformations; that is, $\|UA\|_F = \|A\|_F$, whenever U is an orthonormal matrix.

Example 3.5.7. For $p \in [1, \infty)$ the usual norm on the space $L^p([a, b]; \mathbb{F})$ (see Example 1.1.6(iv)) is

$$\|f\|_{L^p} = \left(\int_a^b |f|^p dx \right)^{1/p}.$$

Similarly, for $p = \infty$ the usual choice of norm on the space $L^\infty([a, b]; \mathbb{F})$ is

$$\|f\|_{L^\infty} = \sup_{x \in [a, b]} |f(x)|.$$

This last example is an especially important one that we use throughout this book. This is sometimes called the *sup norm*.^a

^a*sup* is short for *supremum*.

Definition 3.5.8. For any normed linear space Y and any set X , let $L^\infty(X; Y)$ be the set of all bounded functions from X to Y , that is, functions $f : X \rightarrow Y$ such that the L^∞ -norm $\|f\|_{L^\infty} = \sup_{\mathbf{x} \in X} \|f(\mathbf{x})\|_Y$ is finite.

Proposition 3.5.9. The pair $(L^\infty(X; Y), \|\cdot\|_{L^\infty})$ is a normed vector space.

Proof. The proof is Exercise 3.25. $\square$

3.5.2 Induced Norms on Linear Transformations

In this section, we show how to construct a norm on a set of linear transformations from one normed linear space to another. This allows us to discuss the distance between two linear operators and the convergence of a sequence of linear operators. We often use these norms to prove properties of various linear operators.

Definition 3.5.10. Given two normed linear spaces $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$, let $\mathcal{B}(V; W)$ denote the set of bounded linear transformations, that is, the set of linear maps $T: V \rightarrow W$ for which the quantity

$$\|T\|_{V,W} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} = \sup_{\|\mathbf{x}\|_V=1} \|T(\mathbf{x})\|_W \quad (3.23)$$

is finite. The quantity $\|\cdot\|_{V,W}$ is called the induced norm on $\mathcal{B}(V; W)$, that is, the norm induced by $\|\cdot\|_V$ and $\|\cdot\|_W$. For convenience, if $T: V \rightarrow V$ is a linear operator, then we usually write $\mathcal{B}(V)$ to denote $\mathcal{B}(V; V)$, and we write $\|\cdot\|_V$ to denote²² the induced norm $\|\cdot\|_{V,V}$. We call $\mathcal{B}(V)$ the set of bounded linear operators on V , and we call the induced norm $\|\cdot\|_V$ on $\mathcal{B}(V)$ the operator norm.

Theorem 3.5.11. The set $\mathcal{B}(V; W)$, with operations of vector addition and scalar multiplication defined pointwise, is a vector subspace²³ of $\mathcal{L}(V; W)$, and the pair $(\mathcal{B}(V; W), \|\cdot\|_{V,W})$ is a normed linear space.

Proof. We first prove that (3.23) satisfies the properties of a norm.

(i) Positivity: $\|T\|_{V,W} \geq 0$ by definition. Moreover, if $T = 0$, then $\|T\|_{V,W} = 0$. Conversely, if $T \neq 0$, then $T(\mathbf{x}) \neq \mathbf{0}$ for some $\mathbf{x} \neq \mathbf{0}$. Hence, (3.23) is positive.

(ii) Scale preservation: Note that

$$\|aT\|_{V,W} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|aT(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} = \sup_{\mathbf{x} \neq \mathbf{0}} |a| \frac{\|T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} = |a| \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} = |a| \|T\|_{V,W}.$$

(iii) Triangle inequality:

$$\begin{aligned} \|S + T\|_{V,W} &= \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|S(\mathbf{x}) + T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} \leq \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|S(\mathbf{x})\|_W + \|T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} \\ &\leq \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|S(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} + \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|T(\mathbf{x})\|_W}{\|\mathbf{x}\|_V} = \|S\|_{V,W} + \|T\|_{V,W}. \end{aligned}$$

To prove that $\mathcal{B}(V; W)$ is a vector space, it suffices to show that it is a subspace of $\mathcal{L}(V; W)$. But this follows directly, since scale preservation and the triangle inequality imply that $\mathcal{B}(V; W)$ is closed under vector addition and scalar multiplication. $\square$

Remark 3.5.12. Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed linear spaces. The induced norm of $L \in \mathcal{B}(V; W)$ satisfies $\|L(\mathbf{x})\|_W \leq \|L\|_{V,W} \|\mathbf{x}\|_V$ for all $\mathbf{x} \in V$.

Remark 3.5.13. It is important to note that $\mathcal{B}(V; W) = \mathcal{L}(V; W)$ whenever V and W are both finite dimensional—we prove this in Corollary 3.7.5. But for

²²Even though the notation for the vector norm and the operator norm is the same, the context should make clear which one we are referring to: when what's inside the norm is a vector, it's the vector norm, and when what's inside the norm is an operator, it's the operator norm.

²³See Corollary 2.2.3, which guarantees $\mathcal{L}(V; W)$ is a vector space.

infinite-dimensional vector spaces the space $\mathcal{B}(V; W)$ is usually a proper subspace of $\mathcal{L}(V; W)$. This distinction is important because some of the most useful theorems about linear transformations of finite-dimensional vector spaces extend to $\mathcal{B}(V; W)$ in the infinite-dimensional case but do not hold for $\mathcal{L}(V; W)$.

Theorem 3.5.14. *Let $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$, and $(X, \|\cdot\|_X)$ be normed linear spaces. If $T \in \mathcal{B}(V; W)$ and $S \in \mathcal{B}(W; X)$, then $ST \in \mathcal{B}(V; X)$ and*

$$\|ST\|_{V,X} \leq \|S\|_{W,X} \|T\|_{V,W}. \quad (3.24)$$

In particular, any operator norm $\|\cdot\|_V$ on $\mathcal{B}(V)$ satisfies the submultiplicative property

$$\|ST\|_V \leq \|S\|_V \|T\|_V \quad (3.25)$$

for all S and T in $\mathcal{B}(V)$.

Proof. For all $\mathbf{v} \in V$ we have

$$\|ST\mathbf{v}\|_X \leq \|S\|_{W,X} \|T\mathbf{v}\|_W \leq \|S\|_{W,X} \|T\|_{V,W} \|\mathbf{v}\|_V. \quad \square$$

Definition 3.5.15. *Any norm $\|\cdot\|$ on the finite-dimensional vector space $M_n(\mathbb{F})$ that satisfies the submultiplicative property is called a matrix norm.*

Remark 3.5.16. If $\|\cdot\|$ is an operator norm, then the submultiplicative property shows that for every $n \geq 1$ and every linear operator T , we have $\|T^n\| \leq \|T\|^n$. Moreover, if $\|T\| < 1$, then $\|T^n\| \leq \|T\|^n \rightarrow 0$ as $n \rightarrow \infty$, which implies that $T^n \rightarrow 0$ as $n \rightarrow \infty$ (see Chapter 5). This is a useful observation in many applications, particularly in numerical analysis.

Example 3.5.17. Using the p -norm on both $\mathbb{F}^m$ and $\mathbb{F}^n$ yields the induced matrix norm on $M_{m \times n}(\mathbb{F})$ defined by

$$\|A\|_p = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}, \quad 1 \leq p \leq \infty. \quad (3.26)$$

Unexample 3.5.18. The Frobenius norm (see Example 3.5.6) is not an induced norm, but, as shown in Exercise 4.28, it does satisfy the submultiplicative property $\|AB\|_F \leq \|A\|_F \|B\|_F$.

Example 3.5.19. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space with the usual norm $\|\cdot\|$. If the linear operator $L : V \rightarrow V$ is orthonormal, then $\|L(\mathbf{x})\| = \|\mathbf{x}\|$ for every $\mathbf{x} \in V$, which implies that the induced norm $\|\cdot\|$ on $\mathcal{B}(V)$ satisfies

$\|L\| = 1$. A linear operator that preserves the norm of *every* vector is called an *isometry*. This is a much stronger condition than just saying that $\|L\| = 1$; in fact an isometry preserves both lengths and angles.

3.5.3 Explicit formulas for $\|A\|_1$ and $\|A\|_\infty$

The norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ of a linear operator on a finite-dimensional vector space have very simple descriptions in terms of row and column sums. Although they may be less intuitive than the usual induced 2-norm, they are much simpler to compute and can give us useful bounds for the 2-norm and other norms. This allows us to simplify many problems where more computationally complex norms might seem more natural at first.

Theorem 3.5.20. *Let $A = [a_{ij}] \in M_{m \times n}(\mathbb{F})$. We have the following:*

$$\|A\|_1 = \sup_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|, \quad (3.27)$$

$$\|A\|_\infty = \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|. \quad (3.28)$$

In other words, the 1-norm and the ∞ -norm are, respectively, the largest column and row sums (after taking the modulus of each entry).

Proof. We prove (3.28) and leave (3.27) to the reader (Exercise 3.27). Note that

$$A\mathbf{x} = \begin{bmatrix} \sum_{j=1}^n a_{1j}x_j & \sum_{j=1}^n a_{2j}x_j & \cdots & \sum_{j=1}^n a_{mj}x_j \end{bmatrix}^T.$$

Hence,

$$\|A\mathbf{x}\|_\infty = \sup_{1 \leq i \leq m} \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}||x_j| \leq \left(\sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \right) \|\mathbf{x}\|_\infty.$$

Hence, for all $\mathbf{x} \neq 0$, we have

$$\frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

Taking the supremum of the left side yields

$$\|A\|_\infty \leq \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

It suffices now to prove the reverse inequality over all $\mathbf{x} \in \mathbb{F}^n$. Let k be the row index satisfying

$$\sum_{j=1}^n |a_{kj}| = \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|.$$

Let $\mathbf{x}$ be the vector whose i th entry is 0 if $a_{ki} = 0$ and is $\overline{a_{ki}}/|a_{ki}|$ if $a_{ki} \neq 0$. The only way $\mathbf{x}$ could be $\mathbf{0}$ is if $A = 0$, in which case the theorem is clearly true; so we may assume that at least one of the entries of $\mathbf{x}$ is not zero, and thus $\|\mathbf{x}\|_\infty = 1$. We have

$$\|A\|_\infty \geq \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \sup_{1 \leq i \leq m} \left| \sum_{j=1}^n a_{ij} \frac{\overline{a_{kj}}}{|a_{kj}|} \right| \geq \sum_{j=1}^n |a_{kj}| = \sup_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|. \quad \square$$

Example 3.5.21. If $A = \begin{bmatrix} -1 & 2 \\ 3 & -3 + 4i \end{bmatrix}$, then $\|A\|_1 = 7$ and $\|A\|_\infty = 8$.

Nota Bene 3.5.22. One way to remember that the 1-norm is the largest column sum is to observe that the number one looks like a column. Similarly, the symbol for infinity is more horizontal in shape, corresponding to the fact that the infinity norm is the largest row sum.

3.6 Important Norm Inequalities

The arsenal of the analyst is stocked with inequalities.

—Béla Bollobás

In this section, we examine several important norm inequalities that are fundamental in applied analysis. In particular, we prove the Minkowski and Hölder inequalities. Before doing so, however, we prove Young's inequality, which is one of the most widely used inequalities in all of mathematics.

3.6.1 Young's Inequality

We begin with the following lemma.

Lemma 3.6.1. *If $\frac{1}{p} + \frac{1}{q} = 1$, where $1 < p, q < \infty$ (meaning both $1 < p < \infty$ and $1 < q < \infty$), then for all real $x > 0$ we have*

$$1 \leq \frac{x}{p} + \frac{x^{1-q}}{q}. \quad (3.29)$$

Proof. Let $f(x)$ be the right side of (3.29); that is, $f(x) = \frac{x}{p} + \frac{x^{1-q}}{q}$. Note that $f(1) = 1$. It suffices to show that $x = 1$ is the minimum of $f(x)$. Clearly, $f'(x) = \frac{1}{p} + \frac{x^{-q}(1-q)}{q}$, which simplifies to $f'(x) = \frac{1}{p}(1 - x^{-q})$, since $\frac{1}{p} = \frac{q-1}{q}$. This implies that $x = 1$ is the only critical point. Since $f''(x) = \frac{q}{p}x^{-q-1} > 0$ for all $x > 0$, it follows that f attains its global minimum at $x = 1$. $\square$

Theorem 3.6.2 (Young's Inequality). *If $a, b \geq 0$ and $\frac{1}{p} + \frac{1}{q} = 1$, where $1 < p, q < \infty$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \quad (3.30)$$

Proof. First observe that multiplying $\frac{1}{p} + \frac{1}{q} = 1$ by pq yields $p + q = pq$, and hence $p + q - pq = 0$. By setting $x = \frac{a^{p-1}}{b}$ in (3.29), we have that

$$1 \leq \frac{1}{p} \left(\frac{a^{p-1}}{b} \right) + \frac{1}{q} \left(\frac{a^{p-1}}{b} \right)^{1-q} = \frac{a^{p-1}}{pb} + \frac{a^{p+q-pq-1}}{qb^{1-q}} = \frac{a^{p-1}}{pb} + \frac{b^{q-1}}{qa} = \frac{1}{ab} \left(\frac{a^p}{p} + \frac{b^q}{q} \right).$$

Thus, (3.30) holds. $\square$

Corollary 3.6.3 (Arithmetic-Geometric Mean Inequality). *If $x, y \geq 0$ and $0 \leq \theta \leq 1$, then*

$$x^\theta y^{1-\theta} \leq \theta x + (1-\theta)y. \quad (3.31)$$

In particular, for $\theta = 1/2$, we have $\sqrt{xy} \leq \frac{x+y}{2}$.

Proof. If $\theta = 0$ or $\theta = 1$, then the corollary is trivial. If $\theta \in (0, 1)$, then by Young's inequality with $p = 1/\theta$, $q = 1/(1-\theta)$, $a = x^\theta$, and $b = y^{1-\theta}$, we have

$$x^\theta y^{1-\theta} \leq \theta(x^\theta)^{(1/\theta)} + (1-\theta)(y^{1-\theta})^{1/(1-\theta)} = \theta x + (1-\theta)y. \quad \square$$

3.6.2 Resulting Norm Inequalities

Corollary 3.6.4 (Hölder's Inequality). *If $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ and $\frac{1}{p} + \frac{1}{q} = 1$, where $1 < p, q < \infty$, then*

$$\sum_{k=1}^n |x_k y_k| \leq \left(\sum_{k=1}^n |x_k|^p \right)^{1/p} \left(\sum_{k=1}^n |y_k|^q \right)^{1/q} = \|\mathbf{x}\|_p \|\mathbf{y}\|_q. \quad (3.32)$$

The corresponding result for $p = 1$ and $q = \infty$ also holds:

$$\sum_{k=1}^n |x_k y_k| \leq \left(\sum_{k=1}^n |x_k| \right) \max\{|y_1|, \dots, |y_n|\} = \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty. \quad (3.33)$$

Proof. When $p = 1$ and $q = \infty$, the result is immediate. Since $|y_k| \leq \|\mathbf{y}\|_\infty$, it follows that

$$\sum_{k=1}^n |x_k y_k| \leq \sum_{k=1}^n |x_k| \|\mathbf{y}\|_\infty = \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty.$$

When $1 < p, q < \infty$, Young's inequality gives

$$\sum_{k=1}^n \frac{|x_k|}{\|\mathbf{x}\|_p} \frac{|y_k|}{\|\mathbf{y}\|_q} \leq \sum_{k=1}^n \frac{|x_k|^p}{p \|\mathbf{x}\|_p^p} + \frac{|y_k|^q}{q \|\mathbf{y}\|_q^q} = \frac{\|\mathbf{x}\|_p^p}{p \|\mathbf{x}\|_p^p} + \frac{\|\mathbf{y}\|_q^q}{q \|\mathbf{y}\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1. \quad \square$$

Remark 3.6.5. Assuming the standard inner product on $\mathbb{F}^n$, we have

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q.$$

In particular, when $p = q = 2$, this yields the Cauchy–Schwarz inequality.

Remark 3.6.6. Using Hölder’s inequality on $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{y} = \mathbf{e}_i$ is the standard basis function, we have that $|x_i| \leq \|\mathbf{x}\|_p$.

Corollary 3.6.7 (Minkowski’s Inequality). *If $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ and $p \in (1, \infty)$, then*

$$\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p. \quad (3.34)$$

Minkowski’s inequality is precisely the triangle inequality for the p -norm, thus showing that the p -norm is, indeed, a norm on $\mathbb{F}^n$.

Proof. Choose $q = p/(p-1)$ so that $\frac{1}{p} + \frac{1}{q} = 1$ and $p-1 = \frac{p}{q}$. We have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|_p^p &= \sum_{k=1}^n |x_k + y_k|^p \\ &\leq \sum_{k=1}^n |x_k + y_k|^{p-1} |x_k| + \sum_{k=1}^n |x_k + y_k|^{p-1} |y_k| \\ &\leq \left(\sum_{k=1}^n |x_k + y_k|^p \right)^{1/q} \|\mathbf{x}\|_p + \left(\sum_{k=1}^n |x_k + y_k|^p \right)^{1/q} \|\mathbf{y}\|_p \\ &= \|\mathbf{x} + \mathbf{y}\|_p^{p/q} (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p) \\ &= \|\mathbf{x} + \mathbf{y}\|_p^{p-1} (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p). \end{aligned}$$

Hence, $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. $\square$

Remark 3.6.8. Both Hölder’s and Minkowski’s inequalities can be generalized fairly easily to both ℓ^p and $L^p([a, b]; \mathbb{F})$. Recall that in Example 1.1.6(iv) we prove $L^p([a, b]; \mathbb{F})$ is closed under addition by showing that $\|f + g\|_p \leq 2(\|f\|_p^p + \|g\|_p^p)^{1/p}$. Minkowski’s inequality provides a much sharper bound.

We conclude this section by extending Young’s inequality to give the following two additional norm inequalities.

Corollary 3.6.9. *If $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ and $1 < p, q < \infty$, with $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sum_{i=1}^n |x_i y_i| \leq \frac{1}{p} \sum_{i=1}^n |x_i|^p + \frac{1}{q} \sum_{i=1}^n |y_i|^q = \frac{1}{p} \|\mathbf{x}\|_p^p + \frac{1}{q} \|\mathbf{y}\|_q^q. \quad (3.35)$$

Proof. This follows immediately from (3.30). $\square$

Corollary 3.6.10. *For all $\varepsilon > 0$ and all $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, we have*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sum_{i=1}^n |x_i y_i| \leq \varepsilon \sum_{i=1}^n |x_i|^2 + \frac{1}{4\varepsilon} \sum_{i=1}^n |y_i|^2 = \varepsilon \|\mathbf{x}\|_2^2 + \frac{1}{4\varepsilon} \|\mathbf{y}\|_2^2. \quad (3.36)$$

Proof. This follows immediately from (3.46) in Exercise 3.32. $\square$

3.7 Adjoints

Let A be an $m \times n$ matrix. For the usual inner product (3.2), we have that

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \mathbf{x}^H A\mathbf{y} = (A^H \mathbf{x})^H \mathbf{y} = \langle A^H \mathbf{x}, \mathbf{y} \rangle$$

for any $\mathbf{x} \in \mathbb{F}^m$ and $\mathbf{y} \in \mathbb{F}^n$. In this section, we generalize this property of Hermitian conjugates (see Definition C.1.3) to arbitrary linear transformations and arbitrary inner products. We call the resulting map the *adjoint*.

Before developing the theory of the adjoint, we present the celebrated *Riesz representation theorem*, which states that for each bounded linear transformation $L : V \rightarrow \mathbb{F}$, there exists $\mathbf{w} \in V$ such that $L(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v} \rangle$ for all $\mathbf{v} \in V$. In fact, there is a one-to-one correspondence between the vectors $\mathbf{w}$ and the linear transformations L . In this section, we prove the finite-dimensional version of this result. The infinite-dimensional version is beyond the scope of this text and is typically seen in a standard functional analysis course.

3.7.1 Finite-Dimensional Riesz Representation Theorem

Let $S = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ be a basis of the vector space V . Given a linear transformation $f : V \rightarrow \mathbb{F}$, we can write its matrix representation in the basis S as the $1 \times n$ matrix $[f(\mathbf{x}_1) \ \cdots \ f(\mathbf{x}_n)]$. In other words, if $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{x}_i$, then $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) = [f(\mathbf{x}_1) \ \cdots \ f(\mathbf{x}_n)] \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}^T = \sum_{i=1}^n f(\mathbf{x}_i) a_i.$$

Applying this to $V = \mathbb{F}^n$ with the standard basis $S = [\mathbf{e}_1, \dots, \mathbf{e}_n]$, if $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{e}_i$, then $f(\mathbf{x}) = \sum_{i=1}^n f(\mathbf{e}_i) a_i = \mathbf{y}^H \mathbf{x}$, where $\mathbf{y} = \sum_{i=1}^n \overline{f(\mathbf{e}_i)} \mathbf{e}_i$.

This shows that *every* linear function $f : \mathbb{F}^n \rightarrow \mathbb{F}$ can be written as $f(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$ for some $\mathbf{y} \in \mathbb{F}^n$, where $\langle \cdot, \cdot \rangle$ is the usual inner product. Moreover, we have $\|f(\mathbf{x})\| = |\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{y}\| \|\mathbf{x}\|$, which implies that $\|f\| \leq \|\mathbf{y}\|$. Also, $|\langle \mathbf{y}, \mathbf{x} \rangle| = |f(\mathbf{x})| \leq \|f\| \|\mathbf{x}\|$ for all $\mathbf{x}$, which implies that $\|\mathbf{y}\|^2 \leq \|f\| \|\mathbf{y}\|$, and hence $\|\mathbf{y}\| \leq \|f\|$. Therefore, we have $\|f\| = \|\mathbf{y}\|$. By Corollary 3.3.6 and Remark 3.3.7, these results hold for any finite-dimensional inner product space. We summarize these results in the following theorem.

Theorem 3.7.1 (Finite-Dimensional Riesz Representation Theorem). *Assume that $(V, \langle \cdot, \cdot \rangle)$ is a finite-dimensional inner product space. If L is any linear transformation $L : V \rightarrow \mathbb{F}$, then there exists a unique $\mathbf{y} \in V$ such that $L(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$ for all $\mathbf{x} \in V$. Moreover, we have that $\|L\| = \|\mathbf{y}\| = \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle}$.*

Remark 3.7.2. It is useful to explicitly find the vector $\mathbf{y}$ that the previous theorem promises must exist. Let $[\mathbf{x}_1, \dots, \mathbf{x}_n]$ be an orthonormal basis in V . If $\mathbf{x} \in V$, then we can write $\mathbf{x}$ uniquely as the linear combination $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{x}_i$, where each $a_i = \langle \mathbf{x}_i, \mathbf{x} \rangle$. Hence,

$$L(\mathbf{x}) = \sum_{i=1}^n a_i L(\mathbf{x}_i) = \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{x} \rangle L(\mathbf{x}_i) = \left\langle \sum_{i=1}^n \overline{L(\mathbf{x}_i)} \mathbf{x}_i, \mathbf{x} \right\rangle.$$

If we set $\mathbf{y} = \sum_{i=1}^n \overline{L(\mathbf{x}_i)} \mathbf{x}_i$, then $L(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$ for each $\mathbf{x} \in V$.

Vista 3.7.3. Although the proof of the Riesz representation theorem given above is very simple, it relies on the finite-dimensionality of V . If V is infinite dimensional, then the sum used to define $\mathbf{y}$ becomes an infinite sum, and it is not at all clear that it should converge. Nevertheless, the result can be generalized to infinite dimensions, provided we restrict ourselves to *bounded* linear transformations. The infinite-dimensional Riesz representation theorem is a famous result in functional analysis that has widespread applications in differential equations, probability theory, and optimization, but the proof of the infinite-dimensional case would take us beyond the scope of this book.

Definition 3.7.4. A linear functional is a linear transformation $L : V \rightarrow \mathbb{F}$. If $L \in \mathcal{B}(V; \mathbb{F})$, we say that it is a bounded linear functional.

Corollary 3.7.5. If V is finite dimensional, then every linear functional on V is bounded; that is, $\mathcal{B}(V; \mathbb{F}) = \mathcal{L}(V; \mathbb{F})$. This is a special case of $\mathcal{B}(V; W) = \mathcal{L}(V; W)$, where $W = \mathbb{F}$. (Recall Remark 3.5.13.)

3.7.2 Adjoints

Since every linear functional on a finite-dimensional inner product space V is defined by the inner product with some vector in V , it is natural to ask how such functions and the corresponding inner products change when a linear transformation acts on V . Adjoints answer this question.

Definition 3.7.6. Assume that $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot \rangle_W)$ are inner product spaces and that $L : V \rightarrow W$ is a linear transformation. The adjoint of L is a linear transformation $L^* : W \rightarrow V$ such that

$$\langle \mathbf{w}, L(\mathbf{v}) \rangle_W = \langle L^*(\mathbf{w}), \mathbf{v} \rangle_V \quad \text{for all } \mathbf{v} \in V \text{ and } \mathbf{w} \in W. \quad (3.37)$$

Remark 3.7.7. Theorem 3.7.10 shows that, for finite-dimensional inner product spaces, the adjoint always exists and is unique; so it makes sense to speak of *the* adjoint.

Example 3.7.8. Let $A = [a_{ij}]$ be the matrix representation of $L : \mathbb{F}^m \rightarrow \mathbb{F}^n$ in the standard bases. If $\mathbb{F}^m$ and $\mathbb{F}^n$ have standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y}$, the matrix representation of the adjoint L^* is given by the Hermitian conjugate $A^H = [b_{ji}]$, where $b_{ji} = \bar{a}_{ij}$. This is easy to verify directly:

$$\langle \mathbf{y}, A\mathbf{x} \rangle = \mathbf{y}^H A\mathbf{x} = (A^H \mathbf{y})^H \mathbf{x} = \langle A^H \mathbf{y}, \mathbf{x} \rangle.$$

Example 3.7.9. Let V be the vector space of smooth (infinitely differentiable) functions f on $\mathbb{R}$ that have *compact support*, that is, such that there exists some closed and bounded interval $[a, b]$ such that $f(x) = 0$ if $x \notin [a, b]$. Note that all these functions are Riemann integrable and satisfy $\int_{-\infty}^{\infty} f dx < \infty$. The space V is an inner product space with $\langle f, g \rangle = \int_{-\infty}^{\infty} fg dx$. If we define $L : V \rightarrow V$ by $L[g](x) = g'(x)$, then using integration by parts gives

$$\langle f, L[g] \rangle = \int_{-\infty}^{\infty} f(x)g'(x) dx = - \int_{-\infty}^{\infty} f'(x)g(x) dx = - \langle L[f], g \rangle.$$

So for this inner product and linear operator, we have $L^* = -L$.

Theorem 3.7.10. Assume that $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot \rangle_W)$ are finite-dimensional inner product spaces. If $L : V \rightarrow W$ is a linear transformation, then the adjoint L^* of L exists and is unique.

Proof. To prove existence, define for each $\mathbf{w} \in W$ the linear map $L_{\mathbf{w}} : V \rightarrow \mathbb{F}$ by $L_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{w}, L(\mathbf{v}) \rangle_W$. By the finite-dimensional version of the Riesz representation theorem, there exists a unique $\mathbf{u} \in V$ such that $L_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle_V$ for all $\mathbf{v} \in V$. Thus, we define the map $L^* : W \rightarrow V$ satisfying $L^*(\mathbf{w}) = \mathbf{u}$. Since $\langle \mathbf{w}, L(\mathbf{v}) \rangle_W = \langle L^*(\mathbf{w}), \mathbf{v} \rangle_V$ for all $\mathbf{v} \in V$ and $\mathbf{w} \in W$, we only need to check that L^* is linear. For any $\mathbf{w}_1, \mathbf{w}_2 \in W$ we have

$$\begin{aligned} \langle L^*(a\mathbf{w}_1 + b\mathbf{w}_2), \mathbf{v} \rangle_V &= \langle a\mathbf{w}_1 + b\mathbf{w}_2, L(\mathbf{v}) \rangle_W = \bar{a} \langle \mathbf{w}_1, L(\mathbf{v}) \rangle_W + \bar{b} \langle \mathbf{w}_2, L(\mathbf{v}) \rangle_W \\ &= \bar{a} \langle L^*(\mathbf{w}_1), \mathbf{v} \rangle_V + \bar{b} \langle L^*(\mathbf{w}_2), \mathbf{v} \rangle_V = \langle aL^*(\mathbf{w}_1) + bL^*(\mathbf{w}_2), \mathbf{v} \rangle_V. \end{aligned}$$

Since this holds for all $\mathbf{v} \in V$, we have that $L^*(a\mathbf{w}_1 + b\mathbf{w}_2) = aL^*(\mathbf{w}_1) + bL^*(\mathbf{w}_2)$.

To prove uniqueness, assume that both L_1^* and L_2^* are adjoints of L . We have $\langle (L_1^* - L_2^*)(\mathbf{w}), \mathbf{v} \rangle = 0$ for all $\mathbf{v} \in V$ and $\mathbf{w} \in W$. By Proposition 3.1.16, this implies that $(L_1^* - L_2^*)(\mathbf{w}) = \mathbf{0}$ for all $\mathbf{w} \in W$. Hence, $L_1^* = L_2^*$. $\square$

Vista 3.7.11. As with the Riesz representation theorem, the previous proposition can be extended to bounded linear transformations of infinite-dimensional vector spaces. You should expect to be able to prove this generalization after taking a course in functional analysis.

Proposition 3.7.12. *Let V and W be finite-dimensional inner product spaces. The adjoint has the following properties:*

- (i) *If $S, T \in \mathcal{L}(V; W)$, then $(S + T)^* = S^* + T^*$ and $(\alpha T)^* = \bar{\alpha}T^*$, $\alpha \in \mathbb{F}$.*
- (ii) *If $S \in \mathcal{L}(V; W)$, then $(S^*)^* = S$.*
- (iii) *If $S, T \in \mathcal{L}(V)$, then $(ST)^* = T^*S^*$.*
- (iv) *If $T \in \mathcal{L}(V)$ and T is invertible, then $(T^*)^{-1} = (T^{-1})^*$.*

Proof. The proof is Exercise 3.39. $\square$

3.8 Fundamental Subspaces of a Linear Transformation

In this section we use adjoints to describe four fundamental subspaces associated with a linear operator and prove the fundamental subspaces theorem, which gives some very simple but powerful relations among these spaces.

3.8.1 Orthogonal Complements

Throughout this subsection, let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space.

Definition 3.8.1. *The orthogonal complement of $S \subset V$ is the set*

$$S^\perp = \{\mathbf{y} \in V \mid \langle \mathbf{x}, \mathbf{y} \rangle = 0 \forall \mathbf{x} \in S\}.$$

Example 3.8.2. For a single vector $\mathbf{v} \in V$, the set $\{\mathbf{v}\}^\perp$ is the hyperplane in V defined by $\langle \mathbf{v}, \mathbf{x} \rangle = 0$. Thus, if $V = \mathbb{F}^n$ with the standard inner product and $\mathbf{v} = (v_1, \dots, v_n)$, then $\mathbf{v}^\perp = \{(x_1, \dots, x_n) \mid \bar{v}_1 x_1 + \dots + \bar{v}_n x_n = 0\}$.

Proposition 3.8.3. $S^\perp$ is a subspace of V .

Proof. If $\mathbf{y}_1, \mathbf{y}_2 \in S^\perp$, then for each $\mathbf{x} \in S$, we have $\langle \mathbf{x}, a\mathbf{y}_1 + b\mathbf{y}_2 \rangle = a\langle \mathbf{x}, \mathbf{y}_1 \rangle + b\langle \mathbf{x}, \mathbf{y}_2 \rangle = 0$. Thus, $a\mathbf{y}_1 + b\mathbf{y}_2 \in S^\perp$. $\square$

Remark 3.8.4. For an alternative to the previous proof, recall that the intersection of a collection of subspaces is a subspace (see Proposition 1.2.4). Thus, we have that $S^\perp = \bigcap_{\mathbf{x} \in S} \{\mathbf{x}\}^\perp$. In other words, the hyperplane $\{\mathbf{x}\}^\perp$ is a subspace, and so the intersection of all these subspaces is also a subspace.

Theorem 3.8.5. *If W is a finite-dimensional subspace of V , then $V = W \oplus W^\perp$.*

Proof. If $\mathbf{x} \in W \cap W^\perp$, then $\langle \mathbf{x}, \mathbf{x} \rangle = 0$, which implies that $\mathbf{x} = \mathbf{0}$. Thus, it suffices to prove that $W + W^\perp = V$. This follows from Theorem 3.2.7, since any $\mathbf{v} \in V$ may be written as $\mathbf{v} = \text{proj}_W(\mathbf{v}) + \mathbf{r}$, where $\mathbf{r} = \mathbf{v} - \text{proj}_W(\mathbf{v})$ is orthogonal to W . $\square$

Lemma 3.8.6. *If W is a finite-dimensional subspace of V , then $(W^\perp)^\perp = W$.*

Proof. If $\mathbf{x} \in W$, then $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ for all $\mathbf{y} \in W^\perp$. This implies that $\mathbf{x} \in (W^\perp)^\perp$, and so $W \subset (W^\perp)^\perp$. Suppose now that $\mathbf{x} \in (W^\perp)^\perp$. By Theorem 3.8.5, we can write $\mathbf{x}$ uniquely as $\mathbf{x} = \mathbf{w} + \mathbf{w}_\perp$, where $\mathbf{w} \in W$ and $\mathbf{w}_\perp \in W^\perp$. However, we also have that $\langle \mathbf{w}_\perp, \mathbf{x} \rangle = 0$, which implies that $0 = \langle \mathbf{w}_\perp, \mathbf{w} + \mathbf{w}_\perp \rangle = \langle \mathbf{w}_\perp, \mathbf{w} \rangle + \langle \mathbf{w}_\perp, \mathbf{w}_\perp \rangle = \langle \mathbf{w}_\perp, \mathbf{w}_\perp \rangle = \|\mathbf{w}_\perp\|^2$, and hence $\mathbf{w}_\perp = \mathbf{0}$. This implies $\mathbf{x} \in W$, and thus $(W^\perp)^\perp \subset W$. $\square$

Remark 3.8.7. It is important to note that Theorem 3.8.5 and Lemma 3.8.6 do not hold for all infinite-dimensional subspaces. For example, the space $\mathbb{F}[x]$ of polynomials is a proper subspace of $C([0, 1]; \mathbb{F})$ with the inner product $\langle f, g \rangle = \int_0^1 fg \, dx$, yet it can be shown that the orthogonal complement of $\mathbb{F}[x]$ is the zero vector.

3.8.2 The Fundamental Subspaces

Throughout this subsection, assume that $L : V \rightarrow W$ is a linear transformation from the inner product space $(V, \langle \cdot, \cdot \rangle_V)$ into the inner product space $(W, \langle \cdot, \cdot \rangle_W)$. Assume also that the adjoint of L is L^* .

We now define the four fundamental subspaces.

Definition 3.8.8. *The following are the four fundamental subspaces of L :*

- (i) *The range of L , denoted by $\mathcal{R}(L)$.*
- (ii) *The kernel of L , denoted by $\mathcal{N}(L)$.*
- (iii) *The range of L^* , denoted by $\mathcal{R}(L^*)$.*
- (iv) *The kernel of L^* , denoted by $\mathcal{N}(L^*)$.*

The four fundamental subspaces are depicted in Figure 3.7.

Theorem 3.8.9 (Fundamental Subspaces Theorem). *The following holds for L :*

$$\mathcal{R}(L)^\perp = \mathcal{N}(L^*). \quad (3.38)$$

Moreover, if $\mathcal{R}(L^*)$ is finite dimensional, then

$$\mathcal{N}(L)^\perp = \mathcal{R}(L^*). \quad (3.39)$$

Proof. Note that $\mathbf{w} \in \mathcal{R}(L)^\perp$ if and only if $\langle L(\mathbf{v}), \mathbf{w} \rangle = 0$ for all $\mathbf{v} \in V$, which holds if and only if $\langle \mathbf{v}, L^*\mathbf{w} \rangle = 0$ for all $\mathbf{v} \in V$. This occurs if and only if $L^*\mathbf{w} = \mathbf{0}$, which is equivalent to $\mathbf{w} \in \mathcal{N}(L^*)$. The proof of (3.39) follows from (3.38) and Lemma 3.8.6; see Exercise 3.43. $\square$

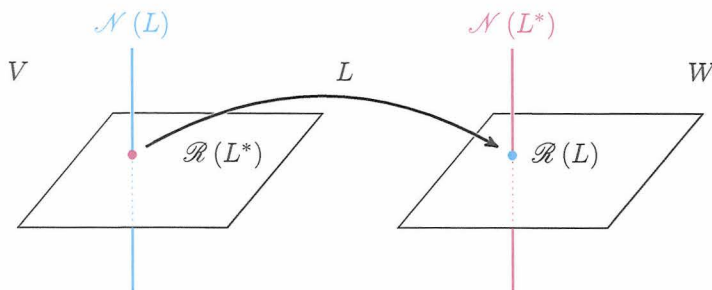


Figure 3.7. The four fundamental subspaces of a linear transformation $L : V \rightarrow W$. Any linear transformation L always maps $\mathcal{R}(L^*)$ (the black plane on the left) isomorphically to $\mathcal{R}(L)$ (the black plane on the right) and sends $\mathcal{N}(L)$ (the vertical blue line on the left) to $\mathbf{0}$ (blue dot on the right) in W . Similarly, L^* maps $\mathcal{R}(L)$ isomorphically to $\mathcal{R}(L^*)$ and sends $\mathcal{N}(L^*)$ (the vertical red line on the right) to $\mathbf{0}$ (the red dot on the left). For an alternative depiction of the fundamental subspaces, see Figure 4.1.

Corollary 3.8.10. If V and W are finite-dimensional vector spaces, then

- (i) $V = \mathcal{N}(L) \oplus \mathcal{R}(L^*)$, and
- (ii) $W = \mathcal{R}(L) \oplus \mathcal{N}(L^*)$.

Moreover, if $n = \dim V$, $m = \dim W$, and $r = \text{rank } L$, the four fundamental subspaces $\mathcal{N}(L)$, $\mathcal{R}(L^*)$, $\mathcal{N}(L^*)$, $\mathcal{R}(L)$ have dimensions $n - r$, r , $m - r$, and r , respectively.

Corollary 3.8.11. Let V and W be finite-dimensional vector spaces. The linear transformation L maps the subspace $\mathcal{R}(L^*)$ bijectively to the subspace $\mathcal{R}(L)$, and L^* maps $\mathcal{R}(L)$ bijectively to $\mathcal{R}(L^*)$.

Proof. Since $\mathcal{N}(L) \cap \mathcal{R}(L^*) = \{\mathbf{0}\}$, the restriction of L to $\mathcal{R}(L^*)$ is injective (see Lemma 2.3.3). Since every $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{x} + \mathbf{r}$ with $\mathbf{r} \in \mathcal{R}(L^*)$ and $\mathbf{x} \in \mathcal{N}(L)$, we have $L(\mathbf{v}) = L(\mathbf{r})$. Thus, $\mathcal{R}(L)$ is the image of L restricted to $\mathcal{R}(L^*)$; that is, L maps $\mathcal{R}(L^*)$ surjectively onto $\mathcal{R}(L)$. Taking adjoints gives the same result for L^* . $\square$

Remark 3.8.12. Let's describe the fundamental subspaces theorem in terms of matrices. Let $V = \mathbb{F}^m$, $W = \mathbb{F}^n$, and let $A = [a_{ij}]$ be the $n \times m$ matrix representing L in the standard bases with the usual inner product. Denote the j th column of A by $\mathbf{a}_j$ and the i th column of A^H by $\mathbf{b}_i$.

The image of a given vector $\mathbf{v} = [v_1 \ \dots \ v_m]^T \in V$ can be written as a linear combination of the columns of A ; that is, $A\mathbf{v} = \sum_{j=1}^m v_j \mathbf{a}_j$. Thus, the range of A is the span of its columns. For this reason, $\mathcal{R}(A)$ is sometimes called the

column space of A , and it has dimension equal to the *rank* of A . Since the adjoint L^* is represented by the matrix A^H , it follows that the range of A^H is the span of its columns.

The null space $\mathcal{N}(A)$ of A is the set of n -tuples $\mathbf{v} = [v_1 \ \dots \ v_n]^T$ in $\mathbb{F}^n$ such that $A\mathbf{v} = \mathbf{0}$, or equivalently

$$\mathbf{0} = A\mathbf{v} = \begin{bmatrix} \mathbf{b}_1^H \\ \mathbf{b}_2^H \\ \vdots \\ \mathbf{b}_m^H \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \langle \mathbf{b}_1, \mathbf{v} \rangle \\ \langle \mathbf{b}_2, \mathbf{v} \rangle \\ \vdots \\ \langle \mathbf{b}_m, \mathbf{v} \rangle \end{bmatrix}.$$

From this we see immediately that $\mathcal{N}(A)$ is orthogonal to the column space $\mathcal{R}(A^H)$, or, in other words, $\mathcal{N}(A)$ is orthogonal to $\mathcal{R}(A^H)$ as described in (3.39). The fundamental subspaces theorem also tells us that the direct sum of these two spaces is all of $V = \mathbb{F}^m$.

Using the same argument for $\mathcal{N}(A^H)$, we get the dual statements that $\mathcal{N}(A^H)$ is orthogonal to the column space $\mathcal{R}(A)$ and that the direct sum of these two spaces is the entire space $W = \mathbb{F}^n$.

Nota Bene 3.8.13. Corollary 3.8.11 shows that $L : \mathcal{R}(L^*) \rightarrow \mathcal{R}(L)$ and $L^* : \mathcal{R}(L) \rightarrow \mathcal{R}(L^*)$ are isomorphisms of vector spaces, but it is important to note that L^* restricted to $\mathcal{R}(L)$ is generally *not* the inverse of L restricted to $\mathcal{R}(L^*)$. Instead, the *Moore–Penrose pseudoinverse* $L^\dagger : W \rightarrow V$ is the inverse of L restricted to $\mathcal{R}(L^*)$ (see Proposition 4.6.2).

Example 3.8.14. Let $A : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be given by

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

The range of A is the span of the columns, which is $\mathcal{R}(A) = \text{span}([1 \ 0]^T)$. This is the x -axis in $\mathbb{R}^2$. The kernel of A is the hyperplane

$$\mathcal{N}(A) = \{[x_1 \ x_2 \ x_3]^T \in \mathbb{R}^3 \mid x_1 + 2x_2 + 3x_3 = 0\},$$

which is orthogonal to $[1 \ 2 \ 3]^T$. Also, $\mathcal{R}(A^H) = \text{span}([1 \ 2 \ 3]^T)$, which by the previous calculation is $\mathcal{N}(A)^\perp$. Similarly, the kernel of A^H is $\mathcal{N}(A^H)$, which is the y -axis in $\mathbb{R}^2$, that is, $\mathcal{R}(A)^\perp$. The adjoint A^H , when restricted to $\mathcal{R}(A)$, is bijective and linear, but it is not quite the inverse of A . To see this, note that

$$AA^H = \begin{bmatrix} 14 & 0 \\ 0 & 0 \end{bmatrix};$$

so when AA^H is restricted to $\mathcal{R}(A) = \text{span}([1 \ 0]^T)$ we have $AA^H|_{\mathcal{R}(A)} = 14I_{\mathcal{R}(A)}$, which is not the identity on $\mathcal{R}(A)$ but is invertible there. Similarly,

$$A^H A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix};$$

so when $A^H A$ is restricted to $\mathcal{R}(A^H) = \text{span}([1 \ 2 \ 3]^T)$ it also acts as multiplication by 14. This can be seen by direct verification: for every $\mathbf{v} \in \text{span}([1 \ 2 \ 3]^T)$ we can write $\mathbf{v} = [t \ 2t \ 3t]^T$ for some $t \in \mathbb{F}$, and we have

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix} \begin{bmatrix} t \\ 2t \\ 3t \end{bmatrix} = 14 \begin{bmatrix} t \\ 2t \\ 3t \end{bmatrix}.$$

So again, $A^H A$ is not the identity on $\mathcal{R}(A^H)$, but is invertible there.

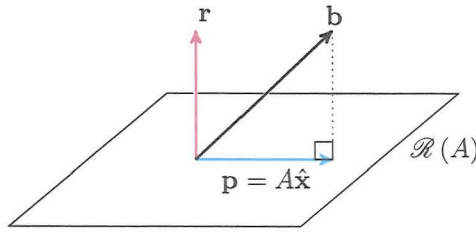


Figure 3.8. Projecting the vector $\mathbf{b}$ onto $\mathcal{R}(A)$ to get $\mathbf{p} = \text{proj}_{\mathcal{R}(A)} \mathbf{b}$ (blue). The best approximate solution to the overdetermined system $A\mathbf{x} = \mathbf{b}$ is the vector $\hat{\mathbf{x}}$ that solves the system $A\hat{\mathbf{x}} = \mathbf{p}$, as described in Section 3.9.1. The error is the norm of the residual $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ (red).

3.9 Least Squares

Many applications involve linear systems that are overdetermined, meaning that there are more equations than unknowns. In this section, we discuss the best approximate solution (called the *least squares solution*) of an overdetermined linear system. This technique is very powerful and is used in virtually every quantitative discipline.

3.9.1 Formulation of Least Squares

Let $A \in M_{m \times n}(\mathbb{F})$. The system $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b} \in \mathcal{R}(A)$. If $\mathbf{b}$ is not in $\mathcal{R}(A)$, we seek the best approximate solution; that is, we want to choose $\hat{\mathbf{x}}$ such that $\|\mathbf{b} - A\hat{\mathbf{x}}\|$ is minimized. By Theorem 3.2.7, this occurs when $A\hat{\mathbf{x}} = \mathbf{p}$, where $\mathbf{p} = \text{proj}_{\mathcal{R}(A)}(\mathbf{b})$. This is depicted in Figure 3.8. Thus, we want to solve

$$A\hat{\mathbf{x}} = \mathbf{p}. \quad (3.40)$$

Since A^H is a bijection when restricted to $\mathcal{R}(A)$ (see Corollary 3.8.11), solving (3.40) is equivalent to solving the equation

$$A^H A \hat{\mathbf{x}} = A^H \mathbf{p}.$$

But $\mathbf{r} = \mathbf{b} - \mathbf{p}$ is orthogonal to $\mathcal{R}(A)$, so by the fundamental subspaces theorem, we have $\mathbf{r} \in \mathcal{N}(A^H)$. Applying the linear transformation A^H , we get

$$A^H \mathbf{b} = A^H (\mathbf{p} + \mathbf{r}) = A^H \mathbf{p}.$$

This tells us that solving (3.40) is equivalent to solving

$$A^H A \hat{\mathbf{x}} = A^H \mathbf{b}, \tag{3.41}$$

which we call the *normal equation*. Any solution $\hat{\mathbf{x}}$ of either (3.41) or (3.40) is called a *least squares solution* of the linear system $A\mathbf{x} = \mathbf{b}$, and it represents the best approximate solution of the linear system in the 2-norm. We summarize this discussion in the following proposition.

Proposition 3.9.1. *For any $A \in M_{m \times n}(\mathbb{F})$ the system $A\mathbf{x} = \mathbf{b}$ has a least squares solution. It is unique if and only if A is injective.*

Why go to the effort of multiplying by A^H ? The main advantage is that the resulting square matrix $A^H A$ is often invertible, as we see in the next lemma. As a side benefit, multiplying by A^H also takes care of projecting to $\mathcal{R}(A)$; that is, it automatically annihilates the part $\mathbf{r}$ of $\mathbf{b}$ that is not in the range of A .

Lemma 3.9.2. *If $A \in M_{m \times n}(\mathbb{F})$ is injective, that is, of rank n , then $A^H A$ is nonsingular.*

Proof. See Exercise 3.46. $\square$

We can now use the invertibility of $A^H A$ to summarize the results of this section as follows.

Theorem 3.9.3. *If $A \in M_{m \times n}(\mathbb{F})$ is injective, that is, of rank n , then the unique least squares solution of the system $A\mathbf{x} = \mathbf{b}$ is given by*

$$\hat{\mathbf{x}} = (A^H A)^{-1} A^H \mathbf{b}. \tag{3.42}$$

Proof. Since A is injective, $A^H A$ is invertible and, when applied to (3.41), gives the unique solution in (3.42). $\square$

Remark 3.9.4. If the matrix A is not injective (not of rank n), then the normal equation (3.41) always has a solution, but the solution is not unique. This generally only occurs in applications when one has collected too few data points, or when variables that one has assumed to be linearly independent are in fact linearly dependent. In situations where A cannot be made injective, there is still a choice of solution that might be considered “best.” We discuss this further when we treat the singular value decomposition in Sections 4.5 and 4.6.

3.9.2 Line and Curve Fitting

Least squares solutions can be used to fit lines or other curves to data. By *fit* we mean that given a family of curves (for example, lines or exponential functions) we want to find the specific curve of that type that most closely matches the data in the sense described above in Section 3.9.1.

For example, suppose we want to fit the line $y = mx + b$ to the data $\{(x_i, y_i)\}_{i=1}^n$. This means we want to find the slope m and y -intercept b so that $mx_i + b = y_i$ for each i . In other words, we want to find the best approximate solution to the overdetermined matrix equation $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} m \\ b \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

The least squares solution is found by solving the normal equation (3.41), which takes the form

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}. \quad (3.43)$$

The matrix $A^T A$ is invertible as long as the x_i terms are not all equal. Simplifying (3.42) yields

$$\begin{bmatrix} m \\ b \end{bmatrix} = \frac{1}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2} \begin{bmatrix} (\sum_{i=1}^n x_i y_i) - n\bar{x}\bar{y} \\ \bar{y}(\sum_{i=1}^n x_i^2) - \bar{x}(\sum_{i=1}^n x_i y_i) \end{bmatrix}, \quad (3.44)$$

where $n\bar{x} = \sum_{i=1}^n x_i$ and $n\bar{y} = \sum_{i=1}^n y_i$.

Example 3.9.5. Consider points $(3.0, 7.3)$, $(4.0, 8.8)$, $(5.0, 11.1)$, and $(6.0, 12.5)$. Using (3.43) or the explicit solution (3.44), we find the line to be $m = 1.79$ and $b = 1.87$; see Figure 3.9(a).

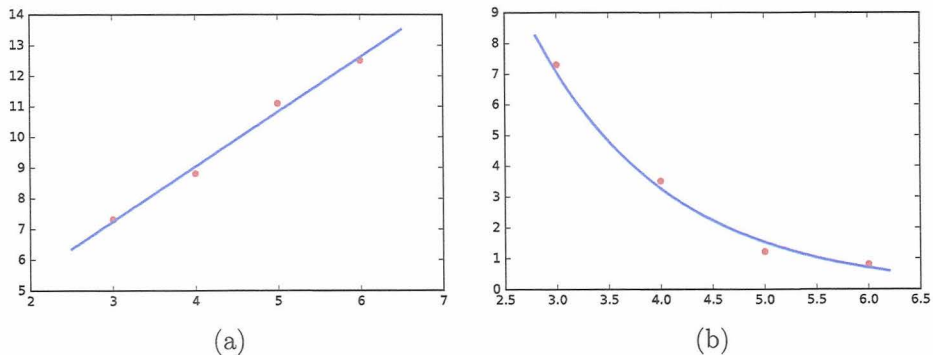


Figure 3.9. Least squares solution fitting (a) a line to four data points and (b) an exponential curve to four points.

Example 3.9.6. Suppose that an amount of radioactive substance is weighed at several points in time, giving n data points $\{(t_i, w_i)\}_{i=1}^n$. We expect radioactive decay to be exponential, so we want to fit the curve $w(t) = ae^{kt}$ to the data. By taking the (natural) log, we can write this as a linear equation $\log w(t) = \log a + kt$. Thus, we want to find the least squares solution for the system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} k \\ \log a \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \log w_1 \\ \log w_2 \\ \vdots \\ \log w_n \end{bmatrix}.$$

If the data points are $(3.0, 7.3)$, $(4.0, 3.5)$, $(5.0, 1.2)$, and $(6.0, 0.8)$, where the first coordinate is measured in years, and the second coordinate is measured in grams, we can use (3.44) to find the exponential parameters to be $k = -0.7703$ and $a = 71.2736$. Thus, the half life in this case is $-(\log 2)/k = 0.8998$ years; see Figure 3.9(b).

Example 3.9.7. Suppose we have a collection of data points $\{(x_i, y_i)\}_{i=1}^n$, and we have reason to believe that they should lie (approximately) on a parabola of the form $y = ax^2 + bx + c$. In this case we want to find the least squares solution to the system $A\mathbf{x} = \mathbf{b}$, with

$$A = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Again the least squares solution is obtained by solving the normal equation $A^H A \mathbf{x} = A^H \mathbf{b}$. The solution is unique if and only if the matrix A has rank 3, which occurs if there are at least three distinct values of x_i in the data set.

Vista 3.9.8. A widely used approach for computing least squares solutions of linear systems is to use the QR decomposition introduced in Section 3.3.3. If $A = QR$ is a QR decomposition of A , then the least squares solution $\hat{\mathbf{x}}$ is found by solving the linear system $R\hat{\mathbf{x}} = Q^H \mathbf{b}$ (see Exercise 3.17). Typically using the QR decomposition takes about twice as long as solving the normal equation directly, but it is more stable.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 3.1. Verify the polarization and parallelogram identities on a real inner product space, with the usual norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ arising from the inner product:

$$(i) \quad \langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} (\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2).$$

$$(ii) \quad \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \frac{1}{2} (\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2).$$

It can be shown that in any normed linear space over $\mathbb{R}$ for which (ii) holds, one can define an inner product by using (i); see [Pro08, Thm. 4.8] for details.

- 3.2. Verify the polarization identity on a complex inner product space, with the usual norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ arising from the inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} (\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 + i\|\mathbf{x} - i\mathbf{y}\|^2 - i\|\mathbf{x} + i\mathbf{y}\|^2).$$

A nice consequence of the polarization identity on a real or complex inner product space is that if two inner products induce the same norm, then the inner products are equal.

- 3.3. Let $\mathbb{R}[x]$ have the inner product

$$\langle f, g \rangle = \int_0^1 f(x)g(x) dx.$$

Using (3.8), find the angle θ between the following sets of vectors:

$$(i) \quad x \text{ and } x^5.$$

$$(ii) \quad x^2 \text{ and } x^4.$$

- 3.4. Let $(V, \langle \cdot, \cdot \rangle)$ be a real inner product space. A linear map $T : V \rightarrow V$ is *angle preserving* if for all nonzero $\mathbf{x}, \mathbf{y} \in V$, we have that

$$\frac{\langle T\mathbf{x}, T\mathbf{y} \rangle}{\|T\mathbf{x}\| \|T\mathbf{y}\|} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (3.45)$$

Prove that T is angle preserving if and only if there exists $a > 0$ such that $\|T\mathbf{x}\| = a\|\mathbf{x}\|$ for all $\mathbf{x} \in V$. Hint: Use Exercise 3.1(i) for one direction. For the other direction, verify that (3.45) implies that T preserves orthogonality, and then write $\mathbf{y}$ as $\mathbf{y} = \text{proj}_{\mathbf{x}} \mathbf{y} + \mathbf{r}$.

- 3.5. Let $V = C([0, 1]; \mathbb{R})$ have the inner product $\langle f, g \rangle = \int_0^1 f(x)g(x) dx$. Find the projection of e^x onto the vector $x - 1$. Hint: Is $x - 1$ a unit vector?
- 3.6. Prove the Cauchy–Schwarz inequality by considering the inequality

$$0 \leq \|\mathbf{x} - \lambda \mathbf{y}\|^2,$$

where $\lambda = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\|\mathbf{y}\|^2}$.

- 3.7. Prove the Cauchy–Schwarz inequality using Corollary 3.2.10. Hint: Consider the orthonormal (singleton) set $\{\frac{\mathbf{x}}{\|\mathbf{x}\|}\}$.
- 3.8. Let V be the inner product space $C([- \pi, \pi]; \mathbb{R})$ with inner product

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)g(t) dt.$$

Let $X = \text{span}(S) \subset V$, where $S = \{\cos(t), \sin(t), \cos(2t), \sin(2t)\}$.

- (i) Prove that S is an orthonormal set.
 - (ii) Compute $\|t\|$.
 - (iii) Compute the projection $\text{proj}_X(\cos(3t))$.
 - (iv) Compute the projection $\text{proj}_X(t)$.
- 3.9. Prove that a rotation (2.17) in $\mathbb{R}^2$ is an orthonormal transformation (with respect to the usual inner product).
- 3.10.  Recall the definition of an orthonormal matrix as given in Definition 3.2.14. Assume the usual inner product on $\mathbb{F}^n$. Prove the following statements:
- (i) The matrix $Q \in M_n(\mathbb{F})$ is an orthonormal matrix if and only if $Q^H Q = Q Q^H = I$.
 - (ii) If $Q \in M_n(\mathbb{F})$ is an orthonormal matrix, then $\|Q\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{F}^n$.
 - (iii) If $Q \in M_n(\mathbb{F})$ is an orthonormal matrix, then so is Q^{-1} .
 - (iv) The columns of an orthonormal matrix $Q \in M_n(\mathbb{F})$ are orthonormal.
 - (v) If $Q \in M_n(\mathbb{F})$ is an orthonormal matrix, then $|\det(Q)| = 1$. Is the converse true?
 - (vi) If $Q_1, Q_2 \in M_n(\mathbb{F})$ are orthonormal matrices, then the product $Q_1 Q_2$ is also an orthonormal matrix.

- 3.11. Describe what happens when we apply the Gram–Schmidt orthonormalization process to a collection of linearly *dependent* vectors.
- 3.12. Apply the Gram–Schmidt orthonormalization process to the set $\{[1, 1]^T, [1, 0]^T\} \subset \mathbb{R}^2$ —this gives a new, orthonormal basis of $\mathbb{R}^2$. Express the vector $2\mathbf{e}_1 + 3\mathbf{e}_2$ in terms of this basis using Theorem 3.2.3.

- 3.13. Apply the Gram–Schmidt orthonormalization process to the set $\{1, x, x^2, x^3\} \subset \mathbb{R}[x]$ with the Chebyshev-1 inner product

$$\langle f, g \rangle = \int_{-1}^1 \frac{fg \, dx}{\sqrt{1-x^2}}.$$

Hint: Recall the trigonometric identity:

$$\cos(u) + \cos(v) = 2 \cos\left(\frac{u+v}{2}\right) \cos\left(\frac{u-v}{2}\right).$$

- 3.14. Prove that for any proper subspace $X \subset V$ of a finite-dimensional inner product space the projection $\text{proj}_X : V \rightarrow X$ is not an orthonormal transformation.
- 3.15. Let

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 4 \\ 1 & -1 \\ 1 & 4 \end{bmatrix}.$$

- (i) Use the Gram–Schmidt method to find the QR decomposition of A .
- (ii) Let $\mathbf{b} = [-1 \ 6 \ 5 \ 7]^T$. Use (i) to solve $A^H A \mathbf{x} = A^H \mathbf{b}$.

- 3.16. Prove the following results about the QR decomposition:

- (i) The QR decomposition is not unique. Hint: Consider matrices of the form QD and $D^{-1}R$, where D is a diagonal matrix.
- (ii) If A is invertible, then there is a unique QR decomposition of A such that R has only positive diagonal elements.

- 3.17.  Let $A \in M_{m \times n}$ have rank $n \leq m$, and let $A = \widehat{Q}\widehat{R}$ be a reduced QR decomposition. Prove that solving the system $A^H A \mathbf{x} = A^H \mathbf{b}$ is equivalent to solving the system $\widehat{R} \mathbf{x} = \widehat{Q}^H \mathbf{b}$.

- 3.18. Let P be the plane $2x + y - z = 0$ in $\mathbb{R}^3$.

- (i) Compute the orthogonal projection of the vector $\mathbf{v} = (1, 1, 1)$ onto the plane P .
- (ii) Compute the matrix representation (in the standard basis) of the reflection through the plane P .
- (iii) Compute the reflection of $\mathbf{v}$ through the plane P .
- (iv) Add $\mathbf{v}$ to the result of (iii). How does this compare to the result of (i)? Explain.

- 3.19. Find the two reflections $H_{\mathbf{v}_1}$ and $H_{\mathbf{v}_2}$ that put the matrix

$$A = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \end{bmatrix}$$

in upper-triangular form; that is, write $H_{\mathbf{v}_2} H_{\mathbf{v}_1} A = R$, where R is upper triangular.

3.20. Prove Proposition 3.4.4.

3.21. Prove Lemma 3.4.7.

3.22* Recall that every rotation around the origin in $\mathbb{R}^2$ can be written in the form (2.17).

(i) Show that the composition of any two Householder reflections in $\mathbb{R}^2$ is a rotation around the origin.

(ii) Prove that any single Householder reflection in $\mathbb{R}^2$ is not a rotation.

3.23. Let $(V, \|\cdot\|)$ be a normed linear space. Prove that $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in V$. Hint: Prove $\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|$ and $\|\mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{y}\|$.

3.24. Let $C([a, b]; \mathbb{F})$ be the vector space of all continuous functions from $[a, b] \subset \mathbb{R}$ to $\mathbb{F}$. Prove that each of the following is a norm on $C([a, b]; \mathbb{F})$:

$$(i) \|f\|_{L^1} = \int_a^b |f(t)| dt.$$

$$(ii) \|f\|_{L^2} = (\int_a^b |f(t)|^2 dt)^{1/2}.$$

$$(iii) \|f\|_{L^\infty} = \sup_{x \in [a, b]} |f(x)|.$$

3.25. Prove Proposition 3.5.9.

3.26.  Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on the vector space X are *topologically equivalent* if there exist constants $0 < m \leq M$ such that

$$m\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq M\|\mathbf{x}\|_a \quad \text{for all } \mathbf{x} \in X.$$

Prove that topological equivalence is an equivalence relation. Then prove that the p -norms for $p = 1, 2, \infty$ on $\mathbb{F}^n$ are topologically equivalent by establishing the following inequalities:

$$(i) \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2.$$

$$(ii) \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty.$$

Hint: Use the Cauchy–Schwarz inequality.

The idea of topological equivalence is especially important in Chapter 5.

3.27. Complete the proof of Theorem 3.5.20 by showing (3.27).

3.28. Let A be an $n \times n$ matrix. Prove that the operator p -norms are topologically equivalent for $p = 1, 2, \infty$ by establishing the following inequalities:

$$(i) \frac{1}{\sqrt{n}}\|A\|_2 \leq \|A\|_1 \leq \sqrt{n}\|A\|_2.$$

$$(ii) \frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\|A\|_\infty.$$

3.29. Take $\mathbb{F}^n$ with the 2-norm, and let the norm on $M_n(\mathbb{F})$ be the corresponding induced norm. Prove that any orthonormal matrix $Q \in M_n(\mathbb{F})$ has $\|Q\| = 1$. For any $\mathbf{x} \in \mathbb{F}^n$, let $R_{\mathbf{x}} : M_n(\mathbb{F}) \rightarrow \mathbb{F}^n$ be the linear transformation $A \mapsto A\mathbf{x}$. Prove that the induced norm of the transformation $R_{\mathbf{x}}$ is equal to $\|\mathbf{x}\|_2$. Hint: First prove $\|R_{\mathbf{x}}\| \leq \|\mathbf{x}\|_2$. Then recall that by Gram–Schmidt, any vector $\mathbf{x}$ with norm $\|\mathbf{x}\|_2 = 1$ is part of an orthonormal basis, and hence is the first column of an orthonormal matrix. Use this to prove equality.

3.30* Let $S \in M_n(\mathbb{F})$ be an invertible matrix. Given any matrix norm $\|\cdot\|$ on M_n , define $\|\cdot\|_S$ by $\|A\|_S = \|SAS^{-1}\|$. Prove that $\|\cdot\|_S$ is a matrix norm on M_n .

- 3.31. Prove that in Young's inequality (3.30), equality holds if and only if $a^p = b^q$.
 3.32. Prove that for every $a, b \geq 0$ and every $\varepsilon > 0$, we have

$$ab \leq \varepsilon a^2 + \frac{b^2}{4\varepsilon}. \quad (3.46)$$

- 3.33. Prove that if $\theta \neq 0, 1$, then equality holds in the arithmetic-geometric mean inequality if and only if $x = y$.
 3.34. Let $(X_1, \|\cdot\|_{X_1}), \dots, (X_n, \|\cdot\|_{X_n})$ be normed linear spaces, and let $X = X_1 \times \cdots \times X_n$ be the Cartesian product. For any $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in X$ define

$$\|\mathbf{x}\|_p = \begin{cases} (\sum_{i=1}^n \|\mathbf{x}_i\|_{X_i}^p)^{1/p} & \text{if } p \in [1, \infty), \\ \sup_i \|\mathbf{x}_i\|_{X_i} & \text{if } p = \infty. \end{cases}$$

For every $p \in [1, \infty]$ prove that $\|\cdot\|_p$ is a norm on X . Hint: Adapt the proof of Minkowski's inequality.

Note that if $X_i = \mathbb{F}$ for every i , then $X = \mathbb{F}^n$ and $\|\cdot\|_p$ is the usual p -norm on $\mathbb{F}^n$.

- 3.35.† Suppose that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $p, q, r \geq 1$ are such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. Prove that

$$\left(\sum_{k=1}^n |x_k y_k|^r \right)^{1/r} \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q.$$

Hint: Note that

$$\frac{1}{\left(\frac{p}{r}\right)} + \frac{1}{\left(\frac{q}{r}\right)} = 1.$$

- 3.36. Use the arithmetic-geometric mean inequality to prove that of all rectangles with a fixed area, the square is the only rectangle with the least perimeter.

- 3.37. Let $V = \mathbb{R}[x; 2]$ be the space of polynomials of degree at most two, which is a subspace of the inner product space $L^2([0, 1]; \mathbb{R})$. Let $L : V \rightarrow \mathbb{R}$ be the linear functional given by $L[p] = p'(1)$. Find the unique $q \in V$ such that $L[p] = \langle q, p \rangle$, as guaranteed by the Riesz representation theorem. Hint: Look at the discussion just before Theorem 3.7.1.

- 3.38. Let $V = \mathbb{F}[x; 2]$, which is a subspace of the inner product space $L^2([0, 1]; \mathbb{R})$. Let D be the derivative operator $D : V \rightarrow V$; that is, $D[p](x) = p'(x)$. Write the matrix representation of D with respect to the power basis $[1, x, x^2]$ of $\mathbb{F}[x; 2]$. Write the matrix representation of the adjoint of D with respect to this basis.

- 3.39.  Prove Proposition 3.7.12.

- 3.40. Let $M_n(\mathbb{F})$ be endowed with the Frobenius inner product (see Example 3.1.7). Any $A \in M_n(\mathbb{F})$ defines a linear operator on $M_n(\mathbb{F})$ by left multiplication: $B \mapsto AB$.

(i) Show that $A^* = A^H$.

(ii) Show that for any $A_1, A_2, A_3 \in M_n(\mathbb{F})$ we have $\langle A_2, A_3 A_1 \rangle = \langle A_2 A_1^*, A_3 \rangle$.
 Hint: Recall $\text{tr}(AB) = \text{tr}(BA)$.

- (iii) Let $A \in M_n(\mathbb{F})$. Define the linear operator $T_A : M_n(\mathbb{F}) \rightarrow M_n(\mathbb{F})$ by $T_A(X) = AX - XA$, and show that $(T_A)^* = T_{A^*}$.

3.41. Let

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 \end{bmatrix}.$$

What are the four fundamental subspaces of A ?

- 3.42. For the linear operator in Exercise 3.38, describe the four fundamental subspaces.
- 3.43. Prove (3.39) in the fundamental subspaces theorem.
- 3.44. Given $A \in M_{m \times n}(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^m$, prove the *Fredholm alternative*: Either $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x} \in \mathbb{F}^n$ or there exists $\mathbf{y} \in \mathcal{N}(A^H)$ such that $\langle \mathbf{y}, \mathbf{b} \rangle \neq 0$.
- 3.45. Consider the vector space $M_n(\mathbb{R})$ with the Frobenius inner product (3.5). Show that $\text{Sym}_n(\mathbb{R})^\perp = \text{Skew}_n(\mathbb{R})$. (See Exercise 1.18 for the definition of Sym and Skew.)

3.46. Prove the following for an $m \times n$ matrix A :

- (i) If $\mathbf{x} \in \mathcal{N}(A^H A)$, then $A\mathbf{x}$ is in both $\mathcal{R}(A)$ and $\mathcal{N}(A^H)$.
- (ii) $\mathcal{N}(A^H A) = \mathcal{N}(A)$.
- (iii) A and $A^H A$ have the same rank.
- (iv) If A has linearly independent columns, then $A^H A$ is nonsingular.

3.47. Assume A is an $m \times n$ matrix of rank n . Let $P = A(A^H A)^{-1}A^H$. Prove the following:

- (i) $P^2 = P$.
- (ii) $P^H = P$.
- (iii) $\text{rank}(P) = n$.

Whenever a linear operator satisfies $P^2 = P$, it is called a *projection*. Projections are treated in detail in Section 12.1.

3.48. Consider the vector space $M_n(\mathbb{R})$ with the Frobenius inner product (3.5).

Let $P(A) = \frac{A + A^T}{2}$ be the map $P : M_n(\mathbb{R}) \rightarrow M_n(\mathbb{R})$. Prove the following:

- (i) P is linear.
- (ii) $P^2 = P$.
- (iii) $P^* = P$ (note that $*$ here means the adjoint with respect to the Frobenius inner product).
- (iv) $\mathcal{N}(P) = \text{Skew}_n(\mathbb{R})$.
- (v) $\mathcal{R}(P) = \text{Sym}_n(\mathbb{R})$.
- (vi) $\|A - P(A)\|_F = \sqrt{\frac{\text{tr}(A^T A) - \text{tr}(A^2)}{2}}$. Here $\|\cdot\|_F$ is the norm with respect to the Frobenius inner product.

Hint: Recall that $\text{tr}(AB) = \text{tr}(BA)$ and $\text{tr}(A) = \text{tr}(A^T)$.

- 3.49. Show that if $A \in M_{m \times n}$, if $\mathbf{b} \in \mathbb{F}^m$, and if $\mathbf{r} = \mathbf{b} - \text{proj}_{\mathcal{R}(A)} \mathbf{b}$, then $\hat{\mathbf{x}} \in \mathbb{F}^n$ is a least squares solution to the linear system $A\mathbf{x} = \mathbf{b}$ if and only if $[\hat{\mathbf{x}}, \mathbf{r}]^T$ is a solution of the equation

$$\begin{bmatrix} A & I \\ 0 & A^H \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

- 3.50. Let $(x_i, y_i)_{i=1}^n$ be a collection of data points that we have reason to believe should lie (roughly) on an ellipse of the form $rx^2 + sy^2 = 1$. We wish to find the least squares approximation for r and s . Write A , $\mathbf{x}$, and $\mathbf{b}$ for the corresponding normal equation in terms of the data x_i and y_i and the unknowns r and s .

Notes

Sources for the infinite-dimensional cases of the results of this chapter include [Pro08, Con90, Rud91]. For more on the QR decomposition and the Householder algorithm, see [TB97, Part II]. For details on the stability of Gaussian elimination, as discussed in Remark 3.3.13, see [TB97, Sect. 22].

4

Spectral Theory

I'm definitely on the spectrum of socially awkward.
—Mayim Bialik

Spectral theory describes how to decouple the domain of a linear operator into a direct sum of minimal components upon which the operator is invariant. Choosing a basis that respects this direct sum results in a corresponding block-diagonal matrix representation. This has powerful consequences in applications, as it allows many problems to be reduced to a series of small individual parts that can be solved independently and more easily. The key tools for constructing this decomposition are *eigenvalues* and *eigenvectors*, which are widely used in many areas of science and engineering and have many important physical interpretations and applications.

For example, they are used to describe the normal modes of vibration in engineered systems such as musical instruments, electrical motors, and even static structures, like bridges and skyscrapers. In quantum mechanics, they describe the possible energy states of an electron or particle; in particular, the atomic orbitals one learns about in chemistry are just eigenvectors of the Hamiltonian operator for a hydrogen atom. Eigenvalues and eigenvectors are fundamental to control theory applications ranging from cruise control on a car to the automated control systems that guide missiles and fly unmanned air vehicles (UAVs) or drones.

Spectral theory is also widely used in the information sciences. For example, eigenvalues and eigenvectors are the key to Google's PageRank algorithm. They can be used in data compression, which in turn is essential for reducing both the complexity and dimensionality in problems like facial recognition, intelligence and personality testing, and machine learning. Eigenvalues and eigenvectors are also useful for decomposing a graph into clusters, which has applications ranging from image segmentation to identifying communities on social media. In short, spectral theory is essential to applied mathematics.

In this chapter, we restrict ourselves to the spectral theory of linear operators on finite-dimensional vector spaces. While it is possible to extend spectral theory to infinite-dimensional spaces, the mathematical sophistication required is beyond the scope of this text and is more suitable for a course in functional analysis.

Most finite-dimensional linear operators have eigenvectors that span the space (hereafter called an *eigenbasis*) where the corresponding matrix representation is diagonal. In other words, a change of basis to the eigenbasis shows that such a matrix operator is similar to a diagonal matrix. That said, not all matrices can be diagonalized, and some can only be made block diagonal. In the first sections of this chapter, we develop this theory and expound on when a matrix can be diagonalized and when we must settle for block diagonalization.

One of the most important and useful results of linear algebra and spectral theory specifically is Schur's lemma, which states that *every* matrix operator is similar to an upper-triangular matrix, and the transition matrix used to perform the similarity transformation is an orthonormal matrix. Any such upper-triangular matrix is called the *Schur form*²⁴ of the operator. The Schur form is well suited to numerical computation, and, as we show in this chapter, it also has real theoretical significance and allows for some very nice proofs of important theorems.

Some matrices have special structure that can make them easier to understand, better behaved, or otherwise more useful than arbitrary matrices. Among the most important of these special matrices are the *normal* matrices, which are characterized by having an orthonormal eigenbasis. This allows a normal matrix to be diagonalized by an orthonormal transition matrix. Among the normal matrices, the *Hermitian* (self-adjoint) matrices are especially important. In this chapter we define and discuss the properties of these and other special classes of matrices.

Finally, the last part of this chapter is devoted to the celebrated singular value decomposition (SVD), which allows any matrix to be separated into parts that describe explicitly how it acts on its fundamental subspaces. This is an essential and very powerful result that you will use over and over, in many different ways.

4.1 Eigenvalues and Eigenvectors

Eigenvectors of an operator are the directions in which the operator acts by simply rescaling;²⁵ the corresponding eigenvalues are the scaling factors. Together the eigenvalues and eigenvectors provide a nice way of understanding the action of the operator; for example, they allow us to write the operator in block-diagonal form.

The eigenvalues of a finite-dimensional operator correspond to the roots of a particular polynomial, called the *characteristic polynomial*. For a given eigenvalue λ , the corresponding eigenvectors generate a subspace called the λ -eigenspace of the operator. When restricted to the eigenspace, the operator just acts as scalar multiplication by λ .

Definition 4.1.1. Let $L : V \rightarrow V$ be a linear operator on a finite-dimensional vector space V over $\mathbb{C}$. A scalar $\lambda \in \mathbb{C}$ is an eigenvalue of L if there exists a nonzero $\mathbf{x} \in V$ such that

$$L(\mathbf{x}) = \lambda\mathbf{x}. \quad (4.1)$$

²⁴Another commonly seen reduced form is the *Jordan canonical form*, which is nonzero only on the diagonal and the superdiagonal. Because the Jordan canonical form has a very nice structure, it is commonly seen in textbooks, but it has only limited use in real-world applications because of inherent computational inaccuracies with floating-point arithmetic. For this reason the Schur form is preferred in most applications. Another important and powerful reduced form is the *spectral decomposition*, which we discuss in Chapter 12.

²⁵This includes reflections, in which eigenvectors are rescaled by a negative number.

Any nonzero $\mathbf{x}$ satisfying (4.1) is called an eigenvector of L corresponding to the eigenvalue λ . For each scalar λ , we define the λ -eigenspace of L as

$$\Sigma_\lambda(L) = \{\mathbf{x} \in V \mid L(\mathbf{x}) = \lambda\mathbf{x}\}. \quad (4.2)$$

The dimension of the λ -eigenspace $\Sigma_\lambda(L)$ is called the geometric multiplicity of λ . If λ is an eigenvalue of L , then $\Sigma_\lambda(L)$ is nontrivial and contains all of its corresponding eigenvectors; otherwise, $\Sigma_\lambda(L) = \{\mathbf{0}\}$. The set of all eigenvalues of L , denoted $\sigma(L)$, is called the spectrum of L . The complement of the spectrum, which is the set of all scalars λ for which $\Sigma_\lambda(L) = \{\mathbf{0}\}$, is called the resolvent set and is denoted $\rho(L)$.

Remark 4.1.2. In addition to all the eigenvectors corresponding to λ , the λ -eigenspace Σ_λ always contains the zero vector $\mathbf{0}$, which is not an eigenvector.

Remark 4.1.3. The definitions of eigenvalues and eigenvectors given above only apply to finite-dimensional operators on complex vector spaces, that is, for matrices with complex entries. For matrices with real entries, we simply think of the real entries as complex numbers. In other words, we use the obvious inclusion $M_n(\mathbb{R}) \subset M_n(\mathbb{C})$ and define eigenvalues and eigenvectors for the corresponding matrices in $M_n(\mathbb{C})$.

The next proposition follows immediately from the definition of $\Sigma_\lambda(L)$.

Proposition 4.1.4. Let $L : V \rightarrow V$ be a linear operator on a finite-dimensional vector space V over $\mathbb{C}$. For any $\lambda \in \mathbb{C}$, the eigenspace $\Sigma_\lambda(L)$ satisfies

$$\Sigma_\lambda(L) = \mathcal{N}(L - \lambda I). \quad (4.3)$$

Nota Bene 4.1.5. Proposition 4.1.4 motivates the traditional method of finding eigenvectors as taught in an elementary linear algebra course. Given that λ is an eigenvalue of the matrix A , you can find the corresponding eigenvectors by solving the linear system $(A - \lambda I)\mathbf{x} = \mathbf{0}$. For a worked example see Example 4.1.12.

Eigenvalues depend only on the linear transformation L and not on a particular choice of basis or matrix representation of L . In other words, they are invariant under similarity transformations. Eigenvectors also depend only on the linear transformation, but their representation changes according to the choice of basis. More precisely, let S be a basis of a finite-dimensional vector space V , and let $L : V \rightarrow V$ be a linear operator with matrix representation A on S . For any eigenvalue λ of L with eigenvector $\mathbf{x}$, we have that (4.1) can be written as

$$A[\mathbf{x}]_S = \lambda[\mathbf{x}]_S.$$

However, if we choose a different basis T on V and denote the corresponding matrix representation of L by B , then the relation between the representations is given,

as always, by the transition matrix C_{TS} . Thus, we have $[\mathbf{x}]_T = C_{TS}[\mathbf{x}]_S$ and $B = C_{TS}A(C_{TS})^{-1}$. This implies that

$$B[\mathbf{x}]_T = C_{TS}A(C_{TS})^{-1}C_{TS}[\mathbf{x}]_S = C_{TS}\lambda[\mathbf{x}]_S = \lambda C_{TS}[\mathbf{x}]_S = \lambda[\mathbf{x}]_T.$$

Conversely, any matrix $A \in M_n(\mathbb{F})$ defines a linear operator on $\mathbb{C}^n$ with the standard basis. We say that $\lambda \in \mathbb{C}$ is an eigenvalue of the matrix A and that the nonzero column vector $\mathbf{x} = [x_1 \ \cdots \ x_n]^T \in \mathbb{C}^n$ is a corresponding eigenvector of the matrix A if $A\mathbf{x} = \lambda\mathbf{x}$. If $B = PAP^{-1}$ is similar to A , then P is the transition matrix into a new basis of $\mathbb{C}^n$, and again we have $BP\mathbf{x} = PAP^{-1}P\mathbf{x} = \lambda P\mathbf{x}$, so λ is also an eigenvalue of the matrix B , and $P\mathbf{x}$ is the corresponding eigenvector.

Remark 4.1.6. The previous discussion shows that the spectral theory of a linear operator on a finite-dimensional vector space and the spectral theory of any associated matrix representation of that operator are essentially the same. But since the representation of the eigenvectors depends on a choice of basis, we will, from this point forth, phrase our study in terms of matrices—that is, in terms of a specific choice of basis.

Unless otherwise stated, we assume throughout the remainder of this chapter that L is a linear operator on a finite-dimensional vector space V and that A is its matrix representation for a given basis S .

Example 4.1.7. Let

$$A = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

We verify that $A\mathbf{x} = -2\mathbf{x}$ and thus conclude that $\lambda = -2$ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$.

Theorem 4.1.8. Let $A \in M_n(\mathbb{F})$ and $\lambda \in \mathbb{C}$. The following are equivalent:

- (i) λ is an eigenvalue of A .
- (ii) There is a nonzero $\mathbf{x}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$.
- (iii) $\Sigma_\lambda(A) \neq \{\mathbf{0}\}$.
- (iv) $\lambda I - A$ is singular.
- (v) $\det(\lambda I - A) = 0$.

Proof. That (i) is equivalent to (ii) follows immediately by subtracting $A\mathbf{x}$ from both sides of the defining equation $A\mathbf{x} = \lambda\mathbf{x}$. That these are equivalent to (iii) follows from Proposition 4.1.4. The fact that these are equivalent to (iv) and (v) follows from Lemma 2.3.3 and Corollary 2.9.6, respectively. $\square$

Definition 4.1.9. Let $A \in M_n(\mathbb{F})$. The polynomial

$$p_A(z) = \det(zI - A) \quad (4.4)$$

is called the characteristic polynomial of A . When it is clear from the context, we sometimes just write $p(z)$ instead of $p_A(z)$.

Remark 4.1.10. According to Theorem 4.1.8, the scalar λ is an eigenvalue of A if and only if $p_A(\lambda) = 0$. Hence, the problem of determining the spectrum of a given matrix is equivalent to locating the roots of its characteristic polynomial.

Proposition 4.1.11. The characteristic polynomial (4.4) is of degree n in z and can be factored over $\mathbb{C}$ as

$$p(z) = \prod_{j=1}^r (z - \lambda_j)^{m_j}, \quad (4.5)$$

where the λ_j are the distinct roots of $p(z)$. The collection of positive integers $(m_j)_{j=1}^r$ satisfy $m_1 + \cdots + m_r = n$ and are called the algebraic multiplicities of the eigenvalues $(\lambda_j)_{j=1}^r$, respectively.²⁶

Proof. Consider the matrix-valued function $B(z) = [b_{ij}] = zI - A$. Since z occurs only on the diagonals of $B(z)$, the degree of an elementary product of $\det(B(z))$ is less than n when the permutation is not the identity and is equal to n when it is the identity. Thus, the characteristic polynomial $p(z)$, which is the sum of the signed elementary products, has degree n . Moreover, by the fundamental theorem of algebra (Theorem 15.3.15), a monic²⁷ degree- n polynomial can be factored into the form (4.5). $\square$

Example 4.1.12. The matrix A from Example 4.1.7 has the characteristic polynomial

$$p(\lambda) = \det(\lambda I - A) = \begin{vmatrix} \lambda - 1 & -3 \\ -4 & \lambda - 2 \end{vmatrix} = \lambda^2 - 3\lambda - 10 = (\lambda - 5)(\lambda + 2).$$

Thus, the spectrum is $\sigma(A) = \{-2, 5\}$.

Once the spectrum is known, the corresponding eigenspaces can be found. Recall from Proposition 4.1.4 that $\Sigma_\lambda(A) = \mathcal{N}(A - \lambda I)$. Hence,

$$\Sigma_5(L) = \mathcal{N}(A - 5I) = \mathcal{N}\left(\begin{bmatrix} -4 & 3 \\ 4 & -3 \end{bmatrix}\right).$$

²⁶Of course, the polynomial $p(z)$ does not necessarily factor into linear terms over $\mathbb{R}$, since $\mathbb{R}$ is not algebraically closed, but it always factors completely over $\mathbb{C}$.

²⁷A polynomial p is *monic* if the coefficient of the term of highest degree is one. For example, $x^2 + 3$ is monic, but $7x^2 + 1$ is not monic.

To find an eigenvector in $\Sigma_5(L)$, just find the values of x and y such that

$$\begin{bmatrix} -4 & 3 \\ 4 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

A short calculation shows that the eigenspace is $\text{span}\{[3 \ 4]^T\}$.

A similar argument gives $\Sigma_{-2}(A) = \mathcal{N}(A + 2I) = \text{span}\{[-1 \ 1]^T\}$. Note that the geometric and algebraic multiplicities of the eigenvalues of A happen to be the same, but this is not true in general.

Example 4.1.13. It is possible to have complex-valued eigenvalues even when the matrix is real-valued. For example, if

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

then $p(\lambda) = \lambda^2 + 1$, and so $\sigma(A) = \{i, -i\}$. By solving for the eigenvectors, we have that $\Sigma_{\pm i}(A) = \mathcal{N}(\pm iI - A) = \text{span}\{[1 \ \mp i]^T\}$. Note again that the geometric and algebraic multiplicities of the eigenvalues of A are the same.

Example 4.1.14. Now consider the matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

Note that $p(\lambda) = (\lambda - 3)^2$, so $\sigma(A) = \{3\}$ and $\mathcal{N}(A - 3I) = \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$. Thus, the algebraic and geometric multiplicities of $\lambda = 3$ are equal to two.

Remark 4.1.15. Despite the previous examples, the geometric and algebraic multiplicities are *not* always the same, as the next example shows.

Example 4.1.16. Consider the matrix

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}.$$

Note that $\sigma(A) = \{3\}$, yet $\mathcal{N}(A - 3I) = \text{span}\{\mathbf{e}_1\}$. Thus, the algebraic multiplicity of $\lambda = 3$ is two, yet the geometric multiplicity is one.

Example 4.1.17. The matrix

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{bmatrix}$$

has characteristic polynomial

$$p(\lambda) = \det(\lambda I - A) = \lambda^3 - \lambda^2 - 5\lambda - 3 = (\lambda - 3)(\lambda + 1)^2,$$

and so $\sigma(A) = \{3, -1\}$. The corresponding eigenspaces are $\Sigma_3 = \mathcal{N}(A - 3I) = \text{span}(\begin{bmatrix} 2 & 1 & 2 \end{bmatrix}^\top)$ and $\Sigma_{-1} = \text{span}(\begin{bmatrix} 2 & -1 & -2 \end{bmatrix}^\top)$. Thus, the algebraic multiplicity of $\lambda = -1$ is two, but the geometric multiplicity is only one.

All finite-dimensional operators (over $\mathbb{C}$) have eigenvalues, but not all operators on infinite-dimensional spaces do. In particular, consider the following example.

Unexample 4.1.18. The definition of eigenvalue extends in an obvious way to operators on infinite-dimensional spaces, but not all operators on infinite-dimensional spaces have eigenvalues. In ℓ^∞ the *right-shift operator* $T : \ell^\infty \rightarrow \ell^\infty$ given by $(a_1, a_2, a_3, \dots) \mapsto (0, a_1, a_2, \dots)$ has no eigenvalue. That is, no choice of $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \ell^\infty$ satisfy the relation $T\mathbf{v} = \lambda\mathbf{v}$.

Application 4.1.19. An important partial differential equation is the Laplace equation, which (in one dimension) takes the form

$$\frac{\partial^2 u}{\partial x^2} = f,$$

where f is a given function. One numerical way to find $u(x)$ in this equation involves solving a linear equation $A\mathbf{x} = \mathbf{b}$, where A is a tridiagonal $n \times n$ matrix in the form

$$A = \begin{bmatrix} b & a & 0 & 0 & \dots & 0 \\ c & b & a & 0 & & \vdots \\ 0 & c & b & a & \ddots & \\ 0 & 0 & c & b & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & a \\ 0 & \dots & 0 & c & b \end{bmatrix}. \quad (4.6)$$

The eigenvalues and eigenvectors of the matrix are also important when solving this problem.

The eigenvalues of A are $\lambda_k = b + 2\sqrt{ac} \cos \omega_{1,k}$ with eigenvectors $\mathbf{x}_k = [\rho \sin \omega_{1,k} \ \cdots \ \rho^n \sin \omega_{n,k}]^T$, where $\omega_{j,k} = \frac{jk\pi}{n+1}$ and $\rho = \left(\frac{c}{a}\right)^{1/2}$. This can be verified by setting $\mathbf{0} = (A - \lambda I)\mathbf{x}_k$, which gives, for each j , that

$$\begin{aligned} 0 &= c\rho^{j-1} \sin \omega_{j-1,k} + (b - \lambda)\rho^j \sin \omega_{j,k} + a\rho^{j+1} \sin \omega_{j+1,k} \\ &= \rho^j (\sqrt{ac} \sin \omega_{j-1,k} + (b - \lambda) \sin \omega_{j,k} + \sqrt{ac} \sin \omega_{j+1,k}) \\ &= \rho^j \sin \omega_{j,k} (2\sqrt{ac} \cos \omega_{1,k} + (b - \lambda)). \end{aligned}$$

Thus, λ_k is an eigenvalue with eigenvector $\mathbf{x}_k$, since these equations all hold when $\lambda = \lambda_k$.

Remark 4.1.20. If A and B are similar matrices, that is, $A = PBP^{-1}$ for some nonsingular P , then A and B define the *same* operator on $\mathbb{F}^n$, but in different bases related by P . Since the determinant and eigenvalues are determined only by a linear operator, and not by its matrix representation, the following proposition is immediate.

Proposition 4.1.21. *If $A, B \in M_n(\mathbb{F})$ are similar matrices, that is, $B = P^{-1}AP$ for some nonsingular matrix P , then the following hold:*

- (i) *A and B have the same characteristic polynomials.*
- (ii) *A and B have the same eigenvalues.*
- (iii) *If λ is an eigenvalue of A and B , then $P : \Sigma_\lambda(B) \rightarrow \Sigma_\lambda(A)$ is an isomorphism, and $\dim \Sigma_\lambda(A) = \dim \Sigma_\lambda(B)$.*

Proof. This follows from Remark 4.1.20, but can also be seen algebraically from the following computation. For any scalar z we have

$$(zI - B) = (zI - P^{-1}AP) = P^{-1}(zI - A)P.$$

This implies

$$\det(zI - B) = \det(P^{-1}(zI - A)P) = \det(P^{-1}) \det(zI - A) \det(P) = \det(zI - A).$$

Moreover, for any $\mathbf{x} \in \Sigma_\lambda(B)$ we know

$$\mathbf{0} = P\mathbf{0} = P(\lambda I - B)\mathbf{x} = PP^{-1}(\lambda I - A)P\mathbf{x} = (\lambda I - A)P\mathbf{x}.$$

So the bijection P maps the subspace $\Sigma_\lambda(B)$ injectively into the subspace $\Sigma_\lambda(A)$. The same argument in reverse shows that P^{-1} maps $\Sigma_\lambda(A)$ into $\Sigma_\lambda(B)$, so they must also be mapped surjectively onto one another, and P is an isomorphism of these subspaces, as required. $\square$

Finally, we conclude with two observations that follow immediately from the results of this section but are nevertheless very useful in many settings.

Proposition 4.1.22. *The diagonal entries of an upper-triangular (or a lower-triangular) matrix are its eigenvalues.*

Proof. See Exercise 4.6. $\square$

Proposition 4.1.23. *A matrix $A \in M_n(\mathbb{F})$ and its transpose A^T have the same characteristic polynomial.*

Proof. See Exercise 4.7. $\square$

4.2 Invariant Subspaces

A subspace is invariant under a given linear operator if all the vectors in the subspace map into vectors in the subspace. More precisely, a subspace is invariant under the linear operator if the domain and codomain of the operator can be restricted to that subspace yielding an operator on the subspace. In this section, we examine the properties of invariant subspaces and show that they can provide a canonical matrix representation for a given linear operator. An important example of an invariant subspace is an eigenspace.

Throughout this section, assume that $L : V \rightarrow V$ is a linear operator on the finite-dimensional vector space V of dimension n .

Definition 4.2.1. *A subspace $W \subset V$ is invariant under L if $L(W) \subset W$. Alternatively, we say that W is L -invariant.*

Nota Bene 4.2.2. Beware that when we say “ W is L -invariant,” it does *not* mean each vector in W is fixed by L . Rather it means that W , as a space, is mapped to itself. Thus, a vector in W could certainly be mapped to a different vector in W , but it will not be sent to a vector outside of W .

Remark 4.2.3. For a vector space V and operator L on V , it is easy to see that $\{0\}$ and V are invariant. But these are not useful—we are really only interested in proper, nontrivial invariant subspaces.

Example 4.2.4. Consider the double-derivative operator L on $\mathbb{F}[x]$ given by $L[p](x) = p''(x)$. The subspace $W = \text{span}\{1, x^2, x^4, x^6, \dots\} \subset \mathbb{F}[x]$ is L -invariant because each basis vector is mapped to W by L ; that is,

$$L[x^{2n}] = 2n(2n-1)x^{2n-2} \in W.$$

You should convince yourself (and prove) that when determining whether a subspace is invariant, it is sufficient to check the basis vectors.

Example 4.2.5. The kernel $\mathcal{N}(L)$ of a linear operator L is always L -invariant because $L(\mathcal{N}(L)) = \{0\} \subset \mathcal{N}(L)$.

Theorem 4.2.6. *If W is an L -invariant subspace of V with basis $S = [\mathbf{s}_1, \dots, \mathbf{s}_k]$, then L restricted to W is a linear operator on W . Let A_{11} denote the matrix representation of L restricted to W , expressed in terms of the basis S . There exists a basis S' of V , containing S , such that the matrix representation A of L on the basis S' is of the form*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}. \quad (4.7)$$

Proof. By Corollary 1.4.5, there exists a set $T = [\mathbf{t}_{k+1}, \mathbf{t}_{k+2}, \dots, \mathbf{t}_n]$ such that $S' = S \cup T$ is a basis for V . Let $A = [a_{ij}]$ be the unique matrix representation of L on S' . Since W is invariant, the map of each basis element of S can be uniquely represented as a linear combination of elements of S . Specifically, we have that $L(\mathbf{s}_j) = \sum_{i=1}^k a_{ij} \mathbf{s}_i$ for $j = 1, \dots, k$. Note that $a_{ij} = 0$ when $i > k$ and $j \leq k$. The image $L(\mathbf{t}_j)$ of each vector in T can be expressed as a linear combination of elements of S' . Thus, we have $L(\mathbf{t}_j) = \sum_{i=1}^k a_{ij} \mathbf{s}_i + \sum_{i=k+1}^n a_{ij} \mathbf{t}_i$ for $j = k+1, \dots, n$. Thus, (4.7) holds, where

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

is the unique matrix representation of L on W on the basis S . $\square$

Example 4.2.7. Let

$$A = \begin{bmatrix} 2 & 3 \\ -1 & -2 \end{bmatrix}$$

be the matrix representation of the linear transformation $L : \mathbb{F}^2 \rightarrow \mathbb{F}^2$ in the standard basis. It is easy to verify that $\text{span}([3 \ -1]^\top)$ is a one-dimensional L -invariant subspace. Choosing any other vector, say $[1 \ 0]^\top$, that is not in the span of $[3 \ -1]^\top$ gives a new basis $\{[3 \ -1]^\top, [1 \ 0]^\top\}$. The corresponding change-of-basis matrices are given by

$$Q = \begin{bmatrix} 3 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad Q^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & 3 \end{bmatrix}.$$

Thus, the representation of L in the new basis is given by

$$Q^{-1}AQ = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix},$$

which is upper triangular.

Remark 4.2.8. The span of an eigenvector is an invariant subspace. Moreover, any eigenspace is also invariant. A finite-dimensional linear operator has a

one-dimensional invariant subspace (corresponding to an eigenvector), but this is not necessarily true for infinite-dimensional spaces; see Exercise 4.9.

Theorem 4.2.9. *Let W_1 and W_2 be L -invariant complementary subspaces of V . If W_1 and W_2 have bases $S_1 = [\mathbf{x}_1, \dots, \mathbf{x}_k]$ and $S_2 = [\mathbf{x}_{k+1}, \dots, \mathbf{x}_n]$, respectively, then the matrix representation A of L on the combined basis $S = S_1 \cup S_2$ is block diagonal; that is,*

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad (4.8)$$

where A_{11} and A_{22} are the matrix representations of L restricted to W_1 and W_2 in the bases S_1 and S_2 , respectively.

Proof. Let $A = [a_{ij}]$ be the matrix representation of L on S . Since W_1 and W_2 are invariant, the map of each basis element of S can be uniquely represented as a linear combination in its respective subspace. Specifically, we have $L(\mathbf{x}_j) = \sum_{i=1}^k a_{ij} \mathbf{x}_i \in W_1$ for $j = 1, \dots, k$ and $L(\mathbf{x}_j) = \sum_{i=k+1}^n a_{ij} \mathbf{x}_i \in W_2$ for $j = k+1, \dots, n$. Thus, the other a_{ij} terms are zero, and (4.8) holds, where

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \quad \text{and} \quad A_{22} = \begin{bmatrix} a_{k+1,k+1} & a_{k+1,k+2} & \cdots & a_{k+1,n} \\ a_{k+2,k+1} & a_{k+2,k+2} & \cdots & a_{k+2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,k+1} & a_{n,k+2} & \cdots & a_{n,n} \end{bmatrix}$$

are the unique matrix representations of L on W_1 and W_2 with respect to the bases S_1 and S_2 , respectively. $\square$

Example 4.2.10. Let $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be a reflection about the line $y = 2x$. It may not be immediately obvious how to write the matrix representation of L in the standard basis S , but Theorem 4.2.9 tells us that if we can find complementary L -invariant subspaces, the matrix representation of L in terms of the bases of those subspaces is diagonal.

A natural choice of complementary, invariant subspaces consists of the line $y = 2x$, which we write as $\text{span}([1 \ 2]^\top)$, and its normal $y = -x/2$, which we can write as $\text{span}([2 \ -1]^\top)$. Let $T = \{[1 \ 2]^\top, [2 \ -1]^\top\}$ be the new basis. Since L leaves $[1 \ 2]^\top$ fixed and takes $[2 \ -1]^\top$ to its negative, we get a very simple matrix representation in this basis:

$$A_{TT} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Changing back to the standard basis via the matrix

$$C_{ST} = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$$

gives the matrix representation

$$C_{ST}A_{TT}(C_{ST})^{-1} = \frac{1}{5} \begin{bmatrix} -3 & 4 \\ 4 & 3 \end{bmatrix}$$

in the standard basis.

Corollary 4.2.11. *Let $W_1, W_2, \dots, W_r$ be a collection of L -invariant subspaces of V . If $V = W_1 \oplus W_2 \oplus \dots \oplus W_r$, where each W_i has the basis S_i , then the unique matrix representation of L on $S = \cup_{i=1}^n S_i$ is of the form*

$$A = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{kk} \end{bmatrix},$$

where each A_{ii} is the matrix representation of L restricted to W_i with the basis S_i .

Proof. This follows by repeated use of Theorem 4.2.9. $\square$

4.3 Diagonalization

If a matrix has a set of eigenvectors that forms a basis, then the matrix is similar to a diagonal matrix. This is one of the most fundamental ideas in linear analysis. In this section we describe how this works. As before, we assume throughout this section that L is a linear operator on a finite-dimensional vector space with matrix representation A , with respect to some given basis.

4.3.1 Simple and Semisimple Matrices

Theorem 4.3.1. *If $\lambda_1, \dots, \lambda_k$ are distinct eigenvalues of L with corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, then these eigenvectors are linearly independent.*

Proof. Suppose $\dim(\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\})) = r < k$. By renumbering we may assume, without loss of generality, that $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r\}$ is linearly independent. Thus, each subsequent vector is a linear combination of the first r vectors. Hence,

$$\mathbf{x}_{r+1} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_r\mathbf{x}_r. \quad (4.9)$$

Applying L yields

$$L\mathbf{x}_{r+1} = a_1L\mathbf{x}_1 + a_2L\mathbf{x}_2 + \cdots + a_rL\mathbf{x}_r,$$

which implies that

$$\lambda_{r+1}\mathbf{x}_{r+1} = a_1\lambda_1\mathbf{x}_1 + a_2\lambda_2\mathbf{x}_2 + \cdots + a_r\lambda_r\mathbf{x}_r. \quad (4.10)$$

Taking (4.10) and subtracting λ_{r+1} times (4.9) yields

$$\mathbf{0} = a_1(\lambda_1 - \lambda_{r+1})\mathbf{x}_1 + a_2(\lambda_2 - \lambda_{r+1})\mathbf{x}_2 + \cdots + a_r(\lambda_r - \lambda_{r+1})\mathbf{x}_r,$$

but since $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r\}$ is linearly independent, each $a_i(\lambda_i - \lambda_{r+1})$ is zero. Because the eigenvalues are distinct, this implies $a_i = 0$ for each i . Hence, $\mathbf{x}_{r+1} = \mathbf{0}$, which is a contradiction (since by definition eigenvectors cannot be $\mathbf{0}$). Therefore, $r = k$. $\square$

Definition 4.3.2. If a set $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbb{F}^n$ of eigenvectors of L forms a basis of $\mathbb{F}^n$, then we say S is an eigenbasis of L .

Example 4.3.3. The matrix

$$A = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

from Examples 4.1.7 and 4.1.12 has distinct eigenvalues, and so the previous theorem shows that the corresponding eigenvectors are linearly independent and thus form an eigenbasis.

Having distinct eigenvalues is a sufficient condition for the existence of an eigenbasis; however, as shown in Example 4.1.14, it is not necessary—the eigenvalues are not distinct, and yet an eigenbasis can be also found. On the other hand, there are cases where an eigenbasis does not exist for a matrix; see, for example, Example 4.1.16.

Definition 4.3.4. An operator L (or the corresponding matrix A) on a finite-dimensional vector space is called

- (i) simple if all of its eigenvalues are distinct;
- (ii) semisimple if there exists an eigenbasis of L .

Corollary 4.3.5. A simple matrix is semisimple.

Proof. If $A \in M_n(\mathbb{F})$ is simple, then there exist n distinct eigenvalues $\lambda_1, \dots, \lambda_n$ with corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Hence, by Theorem 4.3.1, these form a linearly independent set, which is an eigenbasis. $\square$

Definition 4.3.6. The matrix A is diagonalizable if it is similar to a diagonal matrix; that is, there exists a nonsingular matrix P and a diagonal matrix D such that

$$D = P^{-1}AP. \quad (4.11)$$

Theorem 4.3.7. A matrix is diagonalizable if and only if it is semisimple.

Proof. ($\Rightarrow$) If A is diagonalizable, then there exists an invertible matrix P and a diagonal matrix D such that (4.11) holds. If we denote the columns of

P as $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, that is, $P = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]$, and the diagonal elements of D as $\lambda_1, \lambda_2, \dots, \lambda_n$, that is, $D = \text{diag}(\lambda_1, \dots, \lambda_n)$, then it suffices to show that $\lambda_1, \lambda_2, \dots, \lambda_n$ are eigenvalues of A and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are the corresponding eigenvectors, that is, that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$, for each i . Note, however, that this follows from matching up the columns in

$$[A\mathbf{x}_1 \ A\mathbf{x}_2 \ \cdots \ A\mathbf{x}_n] = AP = PD = [\lambda_1\mathbf{x}_1 \ \lambda_2\mathbf{x}_2 \ \cdots \ \lambda_n\mathbf{x}_n].$$

($\Leftarrow$) Assume A is semisimple. Let P be the nonsingular (transition) matrix of eigenvectors $P = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]$. Thus,

$$\begin{aligned} AP &= [A\mathbf{x}_1 \ A\mathbf{x}_2 \ \cdots \ A\mathbf{x}_n] = [\lambda_1\mathbf{x}_1 \ \lambda_2\mathbf{x}_2 \ \cdots \ \lambda_n\mathbf{x}_n] \\ &= [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n] \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) = PD. \end{aligned}$$

It follows that $D = P^{-1}AP$. $\square$

Example 4.3.8. Consider again the matrix

$$A = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

from Examples 4.1.7 and 4.1.12. The eigenvalues are $\sigma(A) = \{-2, 5\}$ with corresponding eigenvectors $[-1 \ 1]^T$ and $[3 \ 4]^T$, respectively. Setting

$$P = \begin{bmatrix} -1 & 3 \\ 1 & 4 \end{bmatrix}, \quad D = \begin{bmatrix} -2 & 0 \\ 0 & 5 \end{bmatrix}, \quad \text{and} \quad P^{-1} = \frac{1}{7} \begin{bmatrix} -4 & 3 \\ 1 & 1 \end{bmatrix},$$

we multiply to find that

$$P^{-1}AP = \frac{1}{7} \begin{bmatrix} -4 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} -1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 5 \end{bmatrix} = D,$$

as expected.

The matrix A represents an operator $L : \mathbb{F}^2 \rightarrow \mathbb{F}^2$ in the standard basis. The previous computation shows that if we change the basis to the eigenbasis $\{[-1 \ 1]^T, [3 \ 4]^T\}$, then the matrix representation of L is D , which is diagonal. This “decouples” the action of the operator L into its action on the two eigenspaces $\Sigma_{-2} = \text{span}([-1 \ 1]^T)$ and $\Sigma_5 = \text{span}([3 \ 4]^T)$.

Example 4.3.9. Not every square matrix is diagonalizable. For example, the matrix

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

in Example 4.1.16 cannot be diagonalized. In this example, $\mathcal{N}(A - 3I) = \text{span}\{\mathbf{e}_1\}$, which is not a basis for $\mathbb{R}^2$, and thus A does not have an eigenbasis.

Diagonalization is useful in many settings. For example, if you diagonalize a matrix, you can compute large powers of it easily since powers of diagonal matrices are trivial to compute, and if $A = P^{-1}DP$, then $A^k = P^{-1}D^kP$ for all $k \in \mathbb{N}$, as shown in the next proposition.

Proposition 4.3.10. *If matrices $A, B \in M_n(\mathbb{F})$ are similar, with $A = P^{-1}BP$, then $A^k = P^{-1}B^kP$ for all $k \in \mathbb{N}$.*

Proof. For $k \in \mathbb{N}$, we have that $A^k = (P^{-1}BP)(P^{-1}BP) \cdots (P^{-1}BP)$, which reduces to $A^k = P^{-1}B^kP$. $\square$

Application 4.3.11. Consider the familiar Fibonacci sequence

$$0, 1, 1, 2, 3, 5, 8, \dots,$$

which can be defined recursively as $F_{n+1} = F_n + F_{n-1}$, where $F_0 = 0$ and $F_1 = 1$. We can also define the sequence in terms of matrices and vectors as follows. Define $\mathbf{v}_k = [F_k \ F_{k-1}]^T$ and observe that

$$\mathbf{v}_{k+1} = A\mathbf{v}_k, \quad \text{where} \quad A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

To find the k th number in the Fibonacci sequence, use the fact that $\mathbf{v}_{k+1} = A\mathbf{v}_k = A^2\mathbf{v}_{k-1} = \cdots = A^k\mathbf{v}_1$.

To calculate A^k , diagonalize A and apply Proposition 4.3.10. A routine calculation shows that the eigenvalues of A are

$$\lambda_1 = \frac{1 + \sqrt{5}}{2} \quad \text{and} \quad \lambda_2 = \frac{1 - \sqrt{5}}{2},$$

and the corresponding eigenvectors are $[\lambda_1 \ 1]^T$ and $[\lambda_2 \ 1]^T$, respectively (The reader should check this!). Since the eigenvectors are linearly independent, they form an eigenbasis, and A can be written as $A = PDP^{-1}$, where

$$P = \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad \text{and} \quad P^{-1} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -\lambda_2 \\ -1 & \lambda_1 \end{bmatrix}.$$

The k th Fibonacci number is the second entry in $\mathbf{v}_{k+1}$, given by

$$\mathbf{v}_{k+1} = PD^kP^{-1}\mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{bmatrix} \begin{bmatrix} 1 & -\lambda_2 \\ -1 & \lambda_1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Multiplying this out gives $F_k = \frac{\lambda_1^k - \lambda_2^k}{\sqrt{5}}$.

Theorem 4.3.12 (Semisimple Spectral Mapping). *If $(\lambda_i)_{i=1}^n$ are the eigenvalues of a semisimple matrix $A \in M_n(\mathbb{F})$ and $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is a polynomial, then $(f(\lambda_i))_{i=1}^n$ are the eigenvalues of $f(A) = a_0I + a_1A + \cdots + a_nA^n$.*

Proof. This is Exercise 4.15. $\square$

Vista 4.3.13. In Section 12.7.1 we show that the semisimple spectral mapping theorem actually holds for all matrices, not just semisimple ones, and it holds for many functions, not just polynomials.

4.3.2 Left Eigenvectors

Recall that an eigenvalue of the matrix $A \in M_n(\mathbb{F})$ is a scalar $\lambda \in \mathbb{F}$ for which $\lambda I - A$ is singular; see Theorem 4.1.8(iv). We have established that an eigenvector of λ is a nonzero element of the kernel $\mathcal{N}(\lambda I - A)$. However, for a given eigenvalue λ , we also have that $\mathbf{x}^\top(\lambda I - A) = \mathbf{0}^\top$ for some nonzero row vector $\mathbf{x}^\top$; see Exercise 4.18. We refer to these row vectors as the *left eigenvectors* of A corresponding to λ . For the remainder of this section, we examine the properties of left eigenvectors.

Definition 4.3.14. Consider a matrix $A \in M_n(\mathbb{F})$. Given an eigenvalue $\lambda \in \mathbb{F}$ we say that the nonzero row vector $\mathbf{x}^\top$ is a *left eigenvector* of A corresponding to λ if

$$\mathbf{x}^\top A = \lambda \mathbf{x}^\top. \quad (4.12)$$

When necessary to avoid confusion, we refer to a regular eigenvector as a right eigenvector.

Remark 4.3.15. It is easy to see by taking the transpose that (4.12) is equivalent to $A^\top \mathbf{x} = \lambda \mathbf{x}$. In other words, given λ , the row vector $\mathbf{x}^\top$ is a left eigenvector of A if and only if $\mathbf{x}$ is a right eigenvector of $A^\top$.

Remark 4.3.16. We define the *left eigenspace* of λ to be the set of row vectors that satisfies (4.12). By appealing to the rank-nullity theorem, we can match the dimensions of the left and right eigenspaces of an eigenvalue λ since it is always true that $\dim \mathcal{N}(\lambda I - A) = \dim \mathcal{N}(\lambda I - A^\top)$.

Remark 4.3.17. Sometimes, for notational convenience, we denote the right eigenvectors by the letter $\mathbf{r}$ and the left eigenvectors by the letter $\ell^\top$.

Example 4.3.18. If

$$B = \begin{bmatrix} -9 & -2 \\ 35 & 8 \end{bmatrix}$$

and $\ell_1^\top = [5 \quad 1]$, then $\ell_1^\top B = -2\ell_1^\top$. Thus, $\lambda = -2$ is a left eigenvalue with $\ell_1^\top$ as a corresponding left eigenvector. Let $\ell_2^\top = [7 \quad 2]$; then $\ell_2^\top B = \ell_2^\top$. Thus, $\lambda = 1$ is a left eigenvalue with $\ell_2^\top$ as a corresponding left eigenvector.

Remark 4.3.19. If a matrix A is semisimple, then it has a basis of right eigenvectors $[\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n]$, which form the transition matrix P used to diagonalize A ; that is, $P^{-1}AP = D$, where $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is the diagonal matrix of eigenvalues (see Theorem 4.3.7). Since $DP^{-1} = P^{-1}A$, it follows for each i that the i th row $\ell_i^\top$ of P^{-1} is a left eigenvector of A with eigenvalue λ_i .

4.4 Schur's Lemma

Schur's lemma states that any square matrix can be transformed via similarity to an upper-triangular matrix and that the similarity transform can be performed with an orthonormal matrix. At first glance, this may seem unimportant, but Schur's lemma is quite powerful, both theoretically and computationally.

Schur's lemma is an important concept in computation. Recall from Proposition 4.1.22 that the eigenvalues of an upper-triangular matrix are given by its diagonal elements. Hence, to find the eigenvalues of a matrix, Schur's lemma provides an alternative to diagonalization. Moreover, the eigenvalues computed with Schur's lemma are generally more accurate than those that are computed by diagonalization.

Given its significance, Schur's lemma is, without question, a theorem in its own right, but for historical reasons it is called a lemma. This is primarily because it nicely sets up the proof of the spectral theorem for Hermitian matrices. Recall that an Hermitian matrix is one that is *self-adjoint*,²⁸ that is, $A = A^H$. The spectral theorem for Hermitian matrices states that all Hermitian matrices are diagonalizable and that their eigenvalues are real. Moreover, there exists an orthonormal eigenbasis that diagonalizes Hermitian matrices.

The most general class of matrices to have orthonormal eigenbases is the class of *normal* matrices, which include Hermitian matrices, skew-Hermitian matrices, and orthonormal matrices. Using Schur's lemma, we show that a matrix is normal if and only if it has an orthonormal eigenbasis.

4.4.1 Schur's Lemma and the Spectral Theorem

Recall from Theorem 3.2.15 that a matrix $Q \in M_n(\mathbb{F})$ is orthonormal if and only if it satisfies $Q^H Q = Q Q^H = I$, which is equivalent to its having orthonormal columns.

Definition 4.4.1. Two matrices A and B are orthonormally similar²⁹ if there exists an orthonormal matrix U such that $B = U^H A U$.

Lemma 4.4.2. If A is Hermitian and orthonormally similar to B , then B is also Hermitian.

Proof. The proof is Exercise 4.20. $\square$

Theorem 4.4.3 (Schur's Lemma). Every matrix $A \in M_n(\mathbb{C})$ is orthonormally similar to an upper-triangular matrix.

Proof. We prove this by induction on n . The $n = 1$ case is trivial. Now assume that the theorem holds for $n = k$, and take $A \in M_{k+1}(\mathbb{C})$. Let λ_1 be an eigenvalue of A with unit eigenvector $\mathbf{w}_1$ (that is, rescale $\mathbf{w}_1$ so that $\|\mathbf{w}_1\|_2 = 1$). Using the Gram-Schmidt algorithm, construct $[\mathbf{w}_2, \dots, \mathbf{w}_{k+1}]$ so that $[\mathbf{w}_1, \dots, \mathbf{w}_{k+1}]$ is an

²⁸These form a very important class of matrices that includes all real, symmetric matrices.

²⁹This is also often called *unitarily similar* in the complex case, and *orthogonally similar* in the real case, but as we have explained before, we prefer to use the name *orthonormal* for both the real and complex situations.

orthonormal set. Setting $U = [\mathbf{w}_1 \cdots \mathbf{w}_{k+1}]$ to be the matrix whose columns are these vectors, we have that

$$U^H A U = \left[\begin{array}{c|ccc} \lambda_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & M & \\ 0 & & & \end{array} \right],$$

where M is a $k \times k$ matrix. By the inductive hypothesis, there exists an orthonormal Q_1 and an upper-triangular T_1 such that $Q_1^H M Q_1 = T_1$. Hence, setting

$$Q = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & Q_1 & \\ 0 & & & \end{array} \right]$$

yields

$$(UQ)^H A (UQ) = Q^H U^H A U Q = \left[\begin{array}{c|ccc} \lambda_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & T_1 & \\ 0 & & & \end{array} \right],$$

which is upper triangular. Finally, since $(UQ)^H (UQ) = Q^H U^H U Q = Q^H I Q = Q^H Q = I$, we have that UQ is orthonormal; see also Corollary 3.2.16. $\square$

Remark 4.4.4. If $B = U^H A U$ is the upper-triangular matrix orthonormally similar to A given by Schur's lemma, it is called the *Schur form* of A . Both A and B correspond to different representations of the same linear operator, and their eigenvalues have the same algebraic and geometric multiplicities. Moreover, the eigenvalues are the diagonal entries of B by Proposition 4.1.22.

Theorem 4.4.5. Let λ be an eigenvalue of an operator T on a finite-dimensional space V . If m_λ is the algebraic multiplicity of λ , then $\dim \Sigma_\lambda(T) \leq m_\lambda$.

Proof. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be a basis of $\Sigma_\lambda(T)$. By the extension theorem (Corollary 1.4.5) we can choose additional vectors $\mathbf{v}_{k+1}, \dots, \mathbf{v}_n$ so that $\mathbf{v}_1, \dots, \mathbf{v}_n$ is a basis of V . The matrix representation A of T in this basis has the form

$$A = \begin{bmatrix} \lambda I_k & * \\ 0 & A_{22} \end{bmatrix},$$

where I_k is the $k \times k$ identity matrix. Thus, the characteristic polynomial $p(z)$ of T satisfies $p(z) = \det(zI - A) = (z - \lambda)^k \det(A_{22})$. But since $p(z)$ factors completely as (4.5), this implies that $k \leq m_\lambda$. $\square$

Corollary 4.4.6. A matrix A is semisimple if and only if $m_\lambda = \dim \Sigma_\lambda(A)$ for each eigenvalue λ .

4.4.2 Spectral Theorem for Hermitian Matrices

Hermitian matrices are self-adjoint complex-valued matrices. These include symmetric real-valued matrices. Self-adjoint operators and Hermitian matrices appear naturally in many physical problems, such as the Schrödinger equation in quantum mechanics and structural models in civil engineering. The spectral theorem tells us that Hermitian matrices have orthonormal eigenbases and that their eigenvalues are always real. We focus primarily on Hermitian matrices and finite-dimensional vector spaces, but the results in this section generalize nicely to self-adjoint operators on infinite-dimensional spaces.

We call the next theorem the *first spectral theorem* because although it is often called just “the spectral theorem,” there is a second, stronger version of the theorem that also should be called “the spectral theorem” (which we have given the clever name the *second spectral theorem*).

Theorem 4.4.7 (First Spectral Theorem). *Every Hermitian matrix A is orthonormally diagonalizable, that is, orthonormally similar to a diagonal matrix. Moreover, the resulting diagonal matrix has only real entries.*

Proof. By Schur's lemma A is orthonormally similar to an upper-triangular matrix T . However, since A is Hermitian, then so is T , by Lemma 4.4.2. This implies that T is diagonal and $\overline{T} = T$; hence, T is real. $\square$

Remark 4.4.8. The converse is also true, since if $A = U^H D U$, with D a real diagonal matrix, then $A^H = (U^H D U)^H = U^H \overline{D} U = A$.

Corollary 4.4.9. *If A is an Hermitian matrix, then it has an orthonormal eigenbasis and the eigenvalues of A are real.*

Proof. Let A be an Hermitian matrix. By the first spectral theorem, there exist an orthonormal matrix U and a diagonal D such that $U^H A U = D$. The eigenbasis is then given by the columns of U , which are orthonormal. The diagonal elements are real since D is also Hermitian (and the diagonal elements of an Hermitian matrix are real). $\square$

Example 4.4.10. Consider the Hermitian matrix

$$A = \begin{bmatrix} 1 & 2i \\ -2i & -2 \end{bmatrix}$$

with characteristic polynomial

$$p(\lambda) = (\lambda - 1)(\lambda + 2) - 4 = \lambda^2 + \lambda - 6 = (\lambda + 3)(\lambda - 2).$$

Thus, the spectrum is $\sigma(A) = \{-3, 2\}$, and so both eigenvalues are real. The corresponding eigenvectors, when scaled to be unit vectors, are

$$\frac{1}{\sqrt{5}} \begin{bmatrix} -i \\ 2 \end{bmatrix} \quad \text{and} \quad \frac{1}{\sqrt{5}} \begin{bmatrix} 2i \\ 1 \end{bmatrix},$$

which form an orthonormal basis for $\mathbb{C}^2$.

Remark 4.4.11. Not every eigenbasis of an Hermitian matrix is orthonormal. First, the eigenvectors need not have unit length. Second, in the case that an Hermitian matrix has an eigenvalue λ of multiplicity two or more, any linearly independent set in the eigenspace of λ can be used to help form an eigenbasis, and that linearly independent set need not be orthogonal. Thus, an orthonormal eigenbasis is a very special choice of eigenbasis.

4.4.3 Normal Matrices

The spectral theorem is extremely powerful, and fortunately it can be generalized to a much larger class of matrices called *normal* matrices. These are matrices that are orthonormally diagonalizable.

Definition 4.4.12. A matrix $A \in M_n(\mathbb{F})$ is normal if $A^H A = A A^H$.

Example 4.4.13. The following are examples of normal matrices:

- (i) Hermitian matrices: $A^H A = A^2 = A A^H$.
- (ii) Skew-Hermitian matrices: $A^H A = -A^2 = A A^H$.
- (iii) Orthonormal matrices: $U^H U = I = U U^H$.

Theorem 4.4.14 (Second Spectral Theorem). A matrix $A \in M_n(\mathbb{F})$ is normal if and only if it is orthonormally diagonalizable.

Proof. Assume that A is normal. By Schur's lemma, there exists an orthonormal matrix U and an upper-triangular matrix T such that $U^H A U = T$. Hence,

$$T^H T = U^H A^H U U^H A U = U^H A^H A U = U^H A A^H U = U^H A U U^H A^H U = T T^H$$

or, in other words, T is also normal. However, we also have that

$$T^H T = \begin{bmatrix} \bar{t}_{11} & 0 & \cdots & 0 \\ \bar{t}_{12} & \bar{t}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{t}_{1n} & \bar{t}_{2n} & \cdots & \bar{t}_{nn} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{bmatrix}$$

and

$$T T^H = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} \bar{t}_{11} & 0 & \cdots & 0 \\ \bar{t}_{12} & \bar{t}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{t}_{1n} & \bar{t}_{2n} & \cdots & \bar{t}_{nn} \end{bmatrix}.$$

Comparing the diagonals of both yields

$$\begin{aligned} |t_{11}|^2 &= |t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2, \\ |t_{12}|^2 + |t_{22}|^2 &= |t_{22}|^2 + |t_{23}|^2 + \cdots + |t_{2n}|^2, \\ &\vdots \\ |t_{1n}|^2 + |t_{2n}|^2 + \cdots + |t_{nn}|^2 &= |t_{nn}|^2. \end{aligned}$$

This implies that $|t_{ij}| = 0$ whenever $i \neq j$. Hence, T is diagonal.

Conversely, if A is orthonormally diagonalizable, there exist an orthonormal matrix U and a diagonal matrix D such that $U^H A U = D$. Thus, $A = U D U^H$. Since $D^H D = D D^H$, we have that $A^H A = U D^H D U^H = U D D^H U^H = A A^H$. Therefore, A is normal. $\square$

Vista 4.4.15. Eigenvalue and eigenvector computations are more prone to error if the matrices involved are not normal. This means that round-off error from floating-point arithmetic and noise in the physical problem can compound into large errors in the final output; see Section 7.5 for details. In Chapter 14, we examine the pseudospectrum, an important tool to better understand the behavior of nonnormal matrices in numerical computation.

4.5 The Singular Value Decomposition

The singular value decomposition (SVD) is one of the most important ideas in applied mathematics and is ubiquitous in science and engineering. The SVD is known by many different names and has several close cousins, each celebrated in different applications and across disciplines. These include the Karhunen–Loève expansion, principal component analysis, factor analysis, empirical orthogonal decomposition, proper orthogonal decomposition, conjoint analysis, the Hotelling transform, latent semantic analysis, and eigenfaces.

The SVD provides orthonormal bases for the four fundamental subspaces as described in Section 3.8. It also gives us a means to approximate a matrix with one of lower rank. Additionally, the SVD allows us to solve least squares problems when the linear systems do not have full column rank.

4.5.1 Positive Definite Matrices

Before we describe the SVD, we must first discuss *positive definite* and *positive semidefinite* matrices, both of which are important for the SVD, but they are also very important in their own right, especially in optimization and statistics.

Definition 4.5.1. A matrix $A \in M_n(\mathbb{F})$ is positive definite, denoted $A > 0$, if it is Hermitian and $\langle \mathbf{x}, A\mathbf{x} \rangle > 0$ for all $\mathbf{x} \neq \mathbf{0}$. It is positive semidefinite, denoted $A \geq 0$, if it is Hermitian and $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all $\mathbf{x}$.

Remark 4.5.2. If A is Hermitian, then it is clear that $\langle \mathbf{x}, A\mathbf{x} \rangle$ is real valued. In particular, we have that

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \langle A^H \mathbf{x}, \mathbf{x} \rangle = \langle A\mathbf{x}, \mathbf{x} \rangle = \overline{\langle \mathbf{x}, A\mathbf{x} \rangle}.$$

Unexample 4.5.3. Consider the matrices

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}.$$

Note that A is not positive definite because it is not Hermitian. To show that B is not positive definite, let $\mathbf{x} = [-1 \ 1]^T$, which gives $\mathbf{x}^H B \mathbf{x} = -4 < 0$.

Example 4.5.4. Consider the matrix

$$C = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

Let $\mathbf{x} = [x_1 \ x_2]^T$ and assume $\mathbf{x} \neq \mathbf{0}$. Thus,

$$\begin{aligned} \mathbf{x}^H C \mathbf{x} &= \bar{x}_1(3x_1 + x_2) + \bar{x}_2(x_1 + 3x_2) \\ &= 3|x_1|^2 + \bar{x}_1 x_2 + x_1 \bar{x}_2 + 3|x_2|^2 \\ &= 2|x_1|^2 + 2|x_2|^2 + |x_1 + x_2|^2 > 0. \end{aligned}$$

Theorem 4.5.5. The Hermitian matrix $A \in M_n(\mathbb{F})$ is positive definite if and only if its spectrum contains only positive eigenvalues. It is positive semidefinite if and only if its spectrum contains only nonnegative eigenvalues.

Proof. ($\Rightarrow$) If λ is an eigenvalue of a positive definite linear operator A with corresponding eigenvector $\mathbf{x}$, then $\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, \lambda \mathbf{x} \rangle = \lambda \|\mathbf{x}\|^2$ is positive. Since $\mathbf{x} \neq \mathbf{0}$, we have $\|\mathbf{x}\| > 0$, and thus λ is positive.

($\Leftarrow$) Let A be Hermitian. If the spectrum $(\lambda_i)_{i=1}^n$ is positive and $(\mathbf{x}_i)_{i=1}^n$ is the corresponding orthonormal eigenbasis, then $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{x}_i$ satisfies

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \left\langle \sum_{i=1}^n a_i \mathbf{x}_i, \sum_{j=1}^n a_j A\mathbf{x}_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \bar{a}_i a_j \lambda_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{i=1}^n |a_i|^2 \lambda_i,$$

which is positive for all $\mathbf{x} \neq \mathbf{0}$.

The same proof works for the positive semidefinite case by replacing every occurrence of the word *positive* with the word *nonnegative*. $\square$

Proposition 4.5.6. If $A \in M_n(\mathbb{F})$ is a positive semidefinite matrix of rank r whose nonzero eigenvalues are $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, then there exists an orthonormal

matrix Q such that $Q^H A Q = \text{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$. The last $n - r$ columns of Q form an orthonormal basis for $\mathcal{N}(A)$, and the first r columns form an orthonormal basis for $\mathcal{N}(A)^\perp$. Letting $D = \text{diag}(\lambda_1, \dots, \lambda_r)$, we have the following block form:

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^H \\ Q_2^H \end{bmatrix}. \quad (4.13)$$

Proof. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be an orthonormal eigenbasis for A , ordered so that the corresponding eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$. The set $\mathbf{x}_{r+1}, \dots, \mathbf{x}_n$ is an orthonormal basis for the kernel $\mathcal{N}(A)$, and $\mathbf{x}_1, \dots, \mathbf{x}_r$ is an orthonormal basis for $\mathcal{N}(A)^\perp$. If Q_1 is defined to be the $n \times r$ matrix with columns equal to $\mathbf{x}_1, \dots, \mathbf{x}_r$, and Q_2 to be the $n \times (n - r)$ matrix with columns equal to $\mathbf{x}_{r+1}, \dots, \mathbf{x}_n$, then (4.13) follows immediately. $\square$

Proposition 4.5.7. *If A is positive semidefinite, there exists a matrix S such that $A = S^H S$. Moreover, if A is positive definite, then the matrix S is nonsingular.*

Proof. Write $A = U D U^H$, where U is orthonormal and $D = \text{diag}(d_1, \dots, d_n) \geq 0$ is diagonal. Let $D^{1/2} = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n}) \geq 0$, so that $D^{1/2} D^{1/2} = D$. Setting $S = D^{1/2} U^H$ gives $S^H S = A$, as required.

If A is positive definite, then each diagonal entry of D is positive, and so is each corresponding diagonal entry of $D^{1/2}$. Thus, D and S are nonsingular. $\square$

Corollary 4.5.8. *Given a positive definite matrix $A \in M_n(\mathbb{F})$, there exists an inner product $\langle \cdot, \cdot \rangle_A$ on $\mathbb{F}^n$ given by $\langle \mathbf{x}, \mathbf{y} \rangle_A = \mathbf{x}^H A \mathbf{y}$.*

Proof. It is clear that $\langle \cdot, \cdot \rangle_A$ is sesquilinear. Since $A > 0$, there exists a nonsingular $S \in M_n(\mathbb{F})$ satisfying $A = S^H S$. Hence, $\langle \mathbf{x}, \mathbf{x} \rangle_A = \mathbf{x}^H A \mathbf{x} = \mathbf{x}^H S^H S \mathbf{x} = \|S \mathbf{x}\|_2^2$, which is positive if and only if $\mathbf{x} \neq \mathbf{0}$. $\square$

Proposition 4.5.9. *If A is an $m \times n$ matrix of rank r , then $A^H A$ is positive semidefinite and has rank r .*

Proof. Note that $(A^H A)^H = A^H A$ and $\langle \mathbf{x}, A^H A \mathbf{x} \rangle = \langle A \mathbf{x}, A \mathbf{x} \rangle = \|A \mathbf{x}\|^2 \geq 0$. Thus, $A^H A$ is positive semidefinite. See Exercise 3.46(iii) to show $\text{rank}(A^H A) = r$. $\square$

4.5.2 The Singular Value Decomposition

The SVD is one of the most important results of this text. For any rank- r matrix $A \in M_{m \times n}(\mathbb{F})$, the SVD gives r positive real numbers $\sigma_1, \dots, \sigma_r$, an orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{F}^n$, and an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_m \in \mathbb{F}^m$ such that A maps the first r vectors $\mathbf{v}_i$ to $\sigma_i \mathbf{u}_i$ and maps the rest of the $\mathbf{v}_i$ to $\mathbf{0}$.

In other words, once we choose the “right” bases for the domain and codomain, any linear transformation of finite-dimensional spaces just maps basis vectors of the

domain to nonnegative real multiples of the basis vectors of the codomain. No matter how complicated a linear transformation may initially appear, in the right bases, it is extremely easy to describe—just a rescaling.

Note that although the SVD involves a diagonal matrix, the SVD is very different from diagonalizing a matrix. When we talk about diagonalizing, we use the same basis for the domain and codomain, but for the SVD we allow them to have different bases. It is this flexibility—the ability to choose different bases for domain and codomain—that allows us to describe the linear transformation in this very simple yet powerful way.

Theorem 4.5.10 (Singular Value Decomposition). *If $A \in M_{m \times n}(\mathbb{F})$ is of rank r , then there exist orthonormal matrices $U \in M_m(\mathbb{F})$ and $V \in M_n(\mathbb{F})$ and an $m \times n$ real diagonal matrix $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r, 0, \dots, 0)$ such that*

$$A = U\Sigma V^H, \quad (4.14)$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ are all positive real numbers. This is called the singular value decomposition (SVD) of A , and the r positive values $\sigma_1, \sigma_2, \dots, \sigma_r$ are the singular values³⁰ of A .

Proof. Let A be an $m \times n$ matrix of rank r . By Proposition 4.5.9, the matrix $A^H A$ is positive semidefinite of rank r . By Proposition 4.5.6, we can find a matrix $V = [V_1 \ V_2]$ with orthonormal columns such that

$$A^H A = [V_1 \ V_2] \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^H \\ V_2^H \end{bmatrix},$$

where $D = \text{diag}(d_1, \dots, d_r)$ and $d_1 \geq d_2 \geq \dots \geq d_r > 0$. Write each d_i as a square $d_i = \sigma_i^2$, and let Σ_1 be the diagonal $r \times r$ block $\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_r)$. In block form, we may write

$$V = [V_1 \ V_2] \quad \text{and} \quad V^H A^H A V = \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix},$$

so we have $D = V_1^H A^H A V_1 = \Sigma_1^2$.

As proved in Exercise 3.46, we have $\mathcal{N}(A^H A) = \mathcal{N}(A)$, which implies that $AV_2 = 0$. Thus, the column vectors of V_2 form an orthonormal basis for $\mathcal{N}(A)$, and the column vectors of V_1 form an orthonormal basis for $\mathcal{N}(A)^\perp = \mathcal{R}(A^H)$.

Define the matrix $U_1 = AV_1 \Sigma_1^{-1}$. It has orthonormal columns because

$$U_1^H U_1 = (\Sigma_1^{-1})^H V_1^H A^H A V_1 \Sigma_1^{-1} = I.$$

Thus, the columns of U_1 form an orthonormal basis for $\mathcal{R}(A)$. Let $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$ be any orthonormal basis for $\mathcal{R}(A)^\perp = \mathcal{N}(A^H)$, and let $U_2 = [\mathbf{u}_{r+1} \ \dots \ \mathbf{u}_m]$ be the matrix with these basis vectors as columns. Setting $U = [U_1 \ U_2]$ yields³¹

$$U\Sigma V^H = [U_1 \ U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^H \\ V_2^H \end{bmatrix} = U_1 \Sigma_1 V_1^H = AV_1 V_1^H. \quad (4.15)$$

³⁰The additional zeros on the diagonal are not considered singular values.

³¹Beware that $V_1 V_1^H \neq I$, despite the fact that $V_1^H V_1 = I$ and $V V^H = I$.

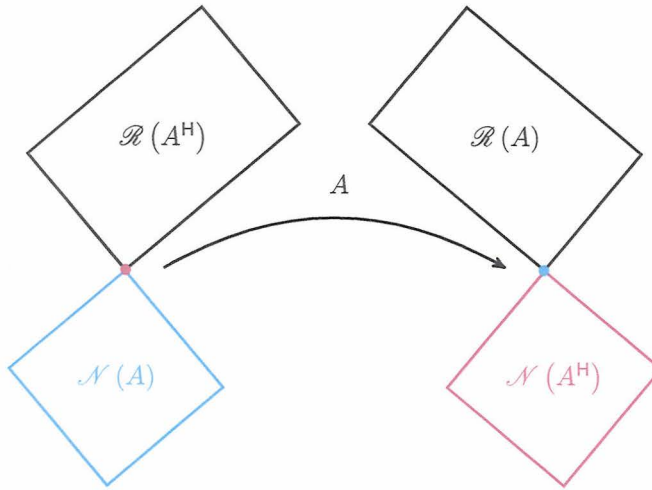


Figure 4.1. This is a representation of the fundamental subspaces theorem (Theorem 3.8.9) that was popularized by Gilbert Strang [Str93]. The transformation A sends $\mathcal{R}(A^H)$ (black rectangle on the left) isomorphically to $\mathcal{R}(A)$ (black rectangle on the right) by sending each basis element $\mathbf{v}_i$ from the SVD to $\sigma_i \mathbf{u}_i \in \mathcal{R}(A)$ if $\mathbf{v}_i \in \mathcal{R}(A^H)$. The remaining $\mathbf{v}_i$ lie in $\mathcal{N}(A)$ (blue square on the left) and are sent to $\mathbf{0}$ (blue dot on the right). Similarly, the transformation A^H maps $\mathcal{R}(A)$ isomorphically to $\mathcal{R}(A^H)$ by sending each $\mathbf{u}_i \in \mathcal{R}(A)$ to $\sigma_i \mathbf{v}_i$. The remaining $\mathbf{u}_i$ lie in $\mathcal{N}(A^H)$ (red square on the right) and are sent to $\mathbf{0}$ (red dot on the left). For an alternative depiction of the fundamental subspaces, see Figure 3.7.

Note that $I = VV^H = V_1V_1^H + V_2V_2^H$, and thus $A = AV_1V_1^H + AV_2V_2^H = AV_1V_1^H$. Therefore, $U\Sigma V^H = A$.

Since the singular values are the positive square roots of the nonzero eigenvalues of $A^H A$, they are uniquely determined by A , and since they are ordered $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$, the matrix Σ is uniquely determined by A . $\square$

Remark 4.5.11. The matrix Σ is unique in the SVD, whereas the matrices U and V are not necessarily unique.

Remark 4.5.12. The SVD gives orthonormal bases for the four fundamental subspaces in the fundamental subspaces theorem (Theorem 3.8.9). Specifically, the first r columns of V form a basis for $\mathcal{R}(A^H)$; the last $n - r$ columns of V form a basis for $\mathcal{N}(A)$; the first r columns of U form a basis for $\mathcal{R}(A)$; and the last $m - r$ columns of U form a basis for $\mathcal{N}(A^H)$. This is visualized in Figure 4.1.

Example 4.5.13. We calculate the SVD of the rank-2 4×3 matrix

$$A = \begin{bmatrix} 2 & 5 & 4 \\ 6 & 3 & 0 \\ 6 & 3 & 0 \\ 2 & 5 & 4 \end{bmatrix}.$$

The first step is to calculate $A^H A$, which is given by

$$\begin{bmatrix} 80 & 56 & 16 \\ 56 & 68 & 40 \\ 16 & 40 & 32 \end{bmatrix}.$$

Since $\sigma(A^H A) = \{0, 36, 144\}$, the singular values are 6 and 12. The right singular vectors, that is, the columns of V , are determined by finding a set of orthonormal eigenvectors of $A^H A$. Specifically, let

$$V = \frac{1}{3} \begin{bmatrix} 2 & -2 & 1 \\ 2 & 1 & -2 \\ 1 & 2 & 2 \end{bmatrix}.$$

The left singular vectors, that is, the columns of U_1 , can be computed by observing that $\mathbf{u}_i = \frac{1}{\sigma_i} A \mathbf{v}_i$ for $i = 1, 2$. The remaining columns of U are calculated by finding unit vectors orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$. We let

$$U = \frac{1}{2} \begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}.$$

Thus, the SVD of A is

$$A = U \Sigma V^H = \left(\frac{1}{2} \begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix} \right) \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left(\frac{1}{3} \begin{bmatrix} 2 & -2 & 1 \\ 2 & 1 & -2 \\ 1 & 2 & 2 \end{bmatrix} \right)^H.$$

Remark 4.5.14. From (4.15) we have that $U \Sigma V^H = U_1 \Sigma_1 V_1^H$. The equation

$$A = U_1 \Sigma_1 V_1^H \quad (4.16)$$

is called the *compact form of the SVD*. The compact form encapsulates all of the necessary information to recalculate the matrix A . Moreover, A can be represented by the *outer-product expansion*

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^H + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^H + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^H, \quad (4.17)$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are column vectors for U_1 and V_1 , respectively, and $\sigma_1, \sigma_2, \dots, \sigma_r$ are the positive singular values.

Example 4.5.15. The compact form of the SVD for the matrix A from Example 4.5.13 is

$$A = U_1 \Sigma_1 V_1^H = \begin{pmatrix} \frac{1}{2} & \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \end{pmatrix} \begin{bmatrix} 12 & 0 \\ 0 & 6 \end{bmatrix} \left(\frac{1}{3} \begin{bmatrix} 2 & -2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix} \right)^H.$$

Corollary 4.5.16 (Polar Decomposition). *If $A \in M_{m \times n}(\mathbb{F})$, with $m \geq n$, then there exists a matrix $Q \in M_{m \times n}(\mathbb{F})$ with orthonormal columns and a positive semidefinite matrix $P \in M_n(\mathbb{F})$ such that $A = QP$. This is called the right polar decomposition of A .*

Proof. From the SVD we have $A = U\Sigma V^H$. Set $Q = UV^H$ and $P = V\Sigma V^H$. Since Σ is positive semidefinite, it follows that P is also positive semidefinite. $\square$

Remark 4.5.17. The polar decomposition is a matrix generalization of writing a complex number in polar form as $re^{i\theta}$; see Appendix B.1. If A is a square matrix, then the determinant of Q lies on the unit circle (see Exercise 3.10(v)). Also, the determinant of P is nonnegative; that is, $\det(A) = \det(Q) \det(P) = re^{i\theta}$, where $\det(P) = r$ and $\det(Q) = e^{i\theta}$.

Remark 4.5.18. Let $A \in M_{m \times n}(\mathbb{F})$, with $m \geq n$. The *left polar decomposition* consists of a matrix $Q \in M_{m \times n}(\mathbb{F})$ with orthonormal columns and a positive semidefinite matrix $P \in M_m(\mathbb{F})$ such that $A = PQ$. This can also be constructed using the SVD.

4.6 Consequences of the SVD

The SVD has many important consequences and applications. In this section we discuss just a few of these.

4.6.1 Least Squares and Moore–Penrose Inverse

The SVD is important in computing least squares solutions when the underlying matrix is not of full column rank; for a reminder of the full-column-rank condition see Theorem 3.9.3. We solve these problems by computing a certain *pseudoinverse*, which behaves like an inverse in many ways.

Consider the linear system $A\mathbf{x} = \mathbf{b}$, where $A \in M_{m \times n}(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^m$. If $\mathbf{b} \in \mathcal{R}(A)$, then the linear system has a solution. If not, then we find the “best” approximate solution in the sense of the 2-norm; see Section 3.9. Recall that the least squares solution $\hat{\mathbf{x}}$ is given by solving the normal equation (3.41) given by

$$A^H A \hat{\mathbf{x}} = A^H \mathbf{b}. \quad (4.18)$$

If A has full column rank, that is, $\text{rank } A = n$, then $A^H A$ is invertible and $\hat{\mathbf{x}} = (A^H A)^{-1} A^H \mathbf{b}$ is the unique least squares solution. If A is not injective, then

there are infinitely many least squares solutions. If $\hat{\mathbf{x}}$ is a particular solution, then any $\hat{\mathbf{x}} + \mathbf{n}$ with $\mathbf{n} \in \mathcal{N}(A)$ also satisfies (4.18). The SVD allows us to find the unique particular solution that is orthogonal to $\mathcal{N}(A)$.

Theorem 4.6.1 (Moore–Penrose Pseudoinverse). *If $A \in M_{m \times n}(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^m$, then there exists a unique $\hat{\mathbf{x}} \in \mathcal{N}(A)^\perp$ satisfying (4.18). Moreover, if $A = U_1 \Sigma_1 V_1^H$ is the compact form of the SVD of A , then $\hat{\mathbf{x}} = A^\dagger \mathbf{b}$, where*

$$A^\dagger = V_1 \Sigma_1^{-1} U_1^H. \quad (4.19)$$

We call $A^\dagger$ the Moore–Penrose pseudoinverse of A .

Proof. If $\hat{\mathbf{x}} = A^\dagger \mathbf{b} = V_1 \Sigma_1^{-1} U_1^H \mathbf{b}$, then $\hat{\mathbf{x}} \in \mathcal{R}(V_1) = \mathcal{R}(A^H) = \mathcal{N}(A)^\perp$ and

$$A^H A \hat{\mathbf{x}} = V_1 \Sigma_1 U_1^H U_1 \Sigma_1 V_1^H V_1 \Sigma_1^{-1} U_1^H \mathbf{b} = V_1 \Sigma_1 U_1^H \mathbf{b} = A^H \mathbf{b}.$$

To prove uniqueness, suppose $\mathbf{v} \in \mathcal{N}(A)^\perp$ is also a solution of the normal equation. Subtracting, we have that $A^H A(\hat{\mathbf{x}} - \mathbf{v}) = \mathbf{0}$, and so $\hat{\mathbf{x}} - \mathbf{v} \in \mathcal{N}(A^H A) = \mathcal{N}(A)$ by Exercise 3.46. Hence, $\hat{\mathbf{x}} - \mathbf{v} \in \mathcal{N}(A)^\perp \cap \mathcal{N}(A) = \{\mathbf{0}\}$. Therefore, $\mathbf{v} = \hat{\mathbf{x}}$. $\square$

Proposition 4.6.2. *If $A \in M_{m \times n}(\mathbb{F})$, then the Moore–Penrose pseudoinverse of A satisfies the following:*

- (i) $AA^\dagger A = A$.
- (ii) $A^\dagger AA^\dagger = A^\dagger$.
- (iii) $(AA^\dagger)^H = AA^\dagger$.
- (iv) $(A^\dagger A)^H = A^\dagger A$.
- (v) $AA^\dagger = \text{proj}_{\mathcal{R}(A)}$ is the orthogonal projection onto $\mathcal{R}(A)$.
- (vi) $A^\dagger A = \text{proj}_{\mathcal{R}(A^H)}$ is the orthogonal projection onto $\mathcal{R}(A^H)$.

Proof. See Exercise 4.38. $\square$

4.6.2 Low-Rank Approximate Behavior

Another important application of the SVD is that it can be used to construct low-rank approximations of a matrix. These are useful for data compression, as well as for many other applications, like facial recognition, because they can greatly reduce the amount of data that must be stored, transmitted, or processed.

The low-rank approximation goes as follows: Consider a matrix $A \in M_{m \times n}(\mathbb{F})$ of rank r . Using the outer product expansion (4.17) of the SVD, and then truncating it to include only the first $s < r$ terms, we define the approximation matrix A_s of A as

$$A_s = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^H + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^H + \cdots + \sigma_s \mathbf{u}_s \mathbf{v}_s^H. \quad (4.20)$$

In this section, we show that A_s has rank s and is the “best” rank- s approximation of A in the sense that the norm of the difference, that is, $\|A - A_s\|$, is minimized against all other $\|A - B\|$ where B has rank s . We make this more precise below.

Theorem 4.6.3 (Schmidt, Mirsky, Eckart–Young).³² *If $A \in M_{m \times n}(\mathbb{C})$ has rank r , then for each $s < r$, we have*

$$\sigma_{s+1} = \inf_{\text{rank}(B)=s} \|A - B\|_2, \quad (4.21)$$

with minimizer

$$B = A_s = \sum_{i=1}^s \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \quad (4.22)$$

where each σ_j is the j th singular value of A and $\mathbf{u}_j$ and $\mathbf{v}_j$ are, respectively, the corresponding columns of U_1 and V_1 in the compact form (4.16) of the singular value decomposition.

Proof. Let $W = [\mathbf{v}_1 \cdots \mathbf{v}_{s+1}]$ be the matrix whose columns are the first $s+1$ right singular vectors of A . For any $B \in M_{m \times n}(\mathbb{C})$ of rank s , Exercise 2.14(i) shows that $\text{rank}(BW) \leq \text{rank}(B) = s$. Thus, by the rank-nullity theorem, we have $\dim \mathcal{N}(BW) = s+1 - \text{rank}(BW) \geq 1$. Hence, there exists $\mathbf{x} \in \mathcal{N}(BW) \subset \mathbb{F}^{s+1}$ satisfying $\|\mathbf{x}\|_2 = 1$. We compute

$$AW\mathbf{x} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^H W\mathbf{x} = \sum_{i=1}^{s+1} \sigma_i x_i \mathbf{u}_i,$$

where $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_{s+1}]^T$. Since $W^H W = I$ (by Theorem 3.2.15), we have that $\|W\mathbf{x}\|_2 = 1$. Thus,

$$\begin{aligned} \|A - B\|_2^2 &= \|A - B\|_2^2 \|W\mathbf{x}\|_2^2 \geq \|(A - B)W\mathbf{x}\|_2^2 = \|AW\mathbf{x}\|_2^2 \\ &= \sum_{i=1}^{s+1} \sigma_i^2 |x_i|^2 \geq \sigma_{s+1}^2 \sum_{i=1}^{s+1} |x_i|^2 = \sigma_{s+1}^2. \end{aligned}$$

This inequality is sharp³³ since

$$\|A - A_s\|_2^2 = \left\| \sum_{i=s+1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^H \right\|_2^2 = \sigma_{s+1}^2,$$

where the last equality is proved in Exercise 4.31(i). $\square$

³²This theorem and its counterpart for the Frobenius norm are often just called the *Eckart–Young theorem*, but there seems to be good evidence that Schmidt and Mirsky discovered these results earlier than Eckart and Young (see [Ste98, pg. 77]), so all four names get attached to these theorems.

³³An inequality is *sharp* if there is at least one case where equality holds. In other words, no stronger inequality could hold.

A version of the previous theorem also holds in the Frobenius norm (as defined in Example 3.5.6).

Theorem 4.6.4 (Schmidt, Mirsky, Eckart–Young). *Using the notation above,*

$$\left(\sum_{j=s+1}^r \sigma_j^2 \right)^{1/2} = \inf_{\text{rank}(B)=s} \|A - B\|_F, \quad (4.23)$$

with minimizer $B = A_s$ given above in (4.22).

Proof. Let $A \in M_{m \times n}(\mathbb{F})$ with SVD $A = U\Sigma V^H$. The invertible change of variable $Z = U^H B V$ (combined with Exercise 4.32(i)) gives

$$\inf_{\text{rank}(B)=s} \|A - B\|_F = \inf_{\text{rank}(Z)=s} \|\Sigma - Z\|_F. \quad (4.24)$$

From Example 3.5.6, we know that the square of the Frobenius norm of a matrix is just the sum of the squares of the entries in the matrix. Hence, if $Z = [z_{ij}]$, then we have

$$\begin{aligned} \|\Sigma - Z\|_F^2 &= \sum_{i=1}^m \sum_{j=1}^n |\Sigma_{ij} - z_{ij}|^2 \\ &= \text{tr}(\Sigma^2) - \text{tr}(\Sigma Z) - \text{tr}(Z^H \Sigma) + \text{tr}(Z^H Z) \\ &= \sum_{i=1}^r \sigma_i^2 - \sum_{i=1}^r \sigma_i (z_{ii} + \overline{z_{ii}}) + \sum_{i=1}^m \sum_{j=1}^n |z_{ij}|^2. \end{aligned}$$

The last expression can only be minimized when $z_{ij} = 0$ for all $i \neq j$ and $z_{ii} = 0$ for $i > r$. Thus, we have that

$$\begin{aligned} \|\Sigma - Z\|_F^2 &= \sum_{i=1}^r \sigma_i^2 - 2 \sum_{i=1}^r \sigma_i \Re(z_{ii}) + \sum_{i=1}^r |z_{ii}|^2 \\ &\geq \sum_{i=1}^r \sigma_i^2 - 2 \sum_{i=1}^r \sigma_i |z_{ii}| + \sum_{i=1}^r |z_{ii}|^2 \\ &= \sum_{i=1}^r (\sigma_i - |z_{ii}|)^2. \end{aligned}$$

Imposing the condition that $\text{rank}(Z) = s$ implies that exactly s of the z_{ii} terms are nonzero. Therefore, the maximum occurs when $z_{ii} = \sigma_i$ for each $1 \leq i \leq s$ and $z_{ii} = 0$ otherwise. This choice knocks out the largest singular values. Hence, the minimizer is (4.22). $\square$

Application 4.6.5 (Data Compression and the SVD). Suppose that you have a large amount of data to transmit or store and only limited bandwidth or storage space. Low-rank approximations give a way to identify and keep only the most important information and discard the rest.

Consider, for example, a grayscale image of dimension 250×250 pixels. The picture can be represented by a 250×250 matrix A , where each entry in the matrix is a number between 0 (black) and 1 (white). This amounts to $250^2 = 62,500$ pixels to transmit or store.

In general the matrix A has rank 250, but the Schmidt, Mirsky, Eckart–Young theorem (Theorem 4.6.4) guarantees that for any $s < \text{rank}(A)$, the matrix

$$A_s = \sum_{i=1}^s \sigma_i \mathbf{u}_i \mathbf{v}_i^H$$

is the best rank- s approximation to A . This matrix can be reconstructed from just the data of the first s singular values and the $2s$ vectors $\mathbf{u}_1, \dots, \mathbf{u}_s$ and $\mathbf{v}_1, \dots, \mathbf{v}_s$ —a total of $501s$ real numbers instead of 62,500. Even with s relatively small, the most important parts of the picture remain. To see an example of the effects of the compression, see Figure 4.2.

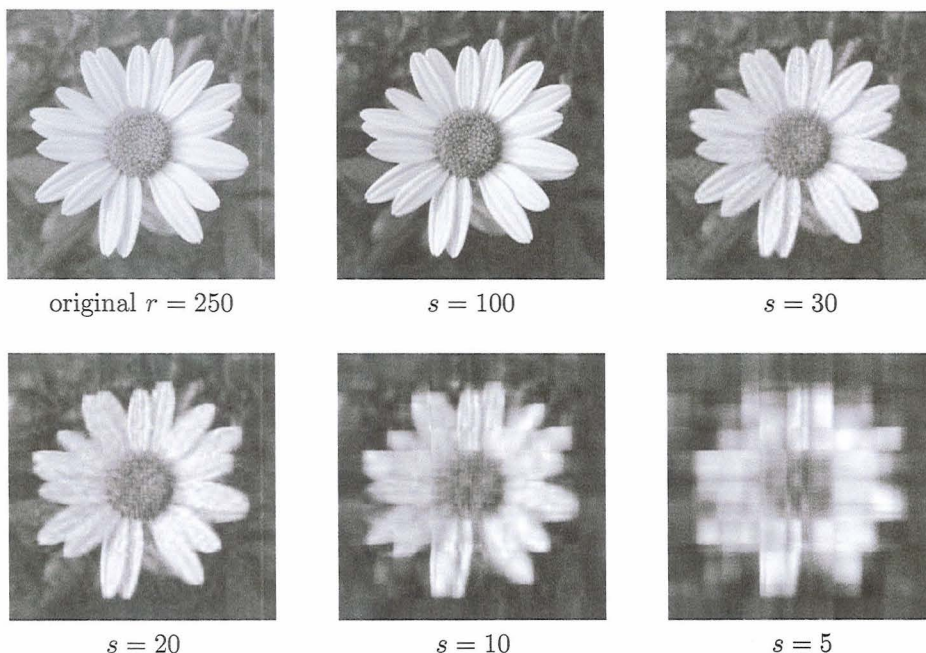


Figure 4.2. An example of image compression as discussed in Application 4.6.5. The original image is 250×250 pixels, and the singular value reconstructions are shown for 100, 30, 20, 10, and 5 singular values. These are the best rank- s approximations of the original image.

An alternative way of formulating the two theorems of Schmidt, Mirsky, and Eckart–Young is that they put an upper bound on the size of a perturbation Δ when $A + \Delta$ has a given rank. So the rank of a small perturbation $A + \Delta$, with Δ sufficiently small, must be at least as large as the rank of A . Put differently,

the matrix Δ has to be sufficiently large in order for $A + \Delta$ to be smaller in rank than A .

Example 4.6.6. If $A = 0 \in M_n(\mathbb{F})$ is the zero matrix, then $A + \varepsilon I$ has rank n for any $\varepsilon \neq 0$. Notice that by adding a small perturbation to the zero matrix, the rank goes from 0 to n . Adding even the smallest matrix to the zero matrix increases the rank of the sum.

Corollary 4.6.7. Let $A \in M_{m \times n}(\mathbb{F})$ have SVD $A = U\Sigma V^H$. If $s < r$, then for any $\Delta \in M_{m \times n}(\mathbb{F})$ satisfying $\text{rank}(A + \Delta) = s$, we have

$$\|\Delta\|_2 \geq \sigma_{s+1} \quad \text{and} \quad \|\Delta\|_F \geq \left(\sum_{k=s+1}^r \sigma_k^2 \right)^{1/2}.$$

Equality holds when $\Delta = -\sum_{i=s+1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^H$.

Proof. The proof is Exercise 4.39. $\square$

4.6.3 Multiplicative Perturbations

We conclude this section by examining multiplicative perturbations. If $A \in M_n(\mathbb{F})$ is invertible, then $I - A\Delta$ has rank zero only if $\Delta = A^{-1}$, which has 2-norm equal to $\|A^{-1}\|_2 = \sigma_n^{-1}$; see Exercise 4.31(ii). To force $I - A\Delta$ to have rank less than n (but not necessarily 0) does not require Δ to be so large, but we still have a lower bound for Δ .

Theorem 4.6.8. Let $A \in M_{m \times n}(\mathbb{F})$ have SVD $A = U\Sigma V^H$. The infimum of $\|\Delta\|$ such that $\text{rank}(I - A\Delta) < m$ is σ_1^{-1} , with

$$\sigma_1^{-1} = \inf_{\text{rank}(I - A\Delta) < m} \|\Delta\|_2, \quad (4.25)$$

and with minimizer $\Delta^* = \sigma_1^{-1} \mathbf{v}_1 \mathbf{u}_1^H$.

Proof. To make sense of the expression $I - A\Delta$, we must have $I \in M_{m \times m}(\mathbb{F})$ and $\Delta \in M_{n \times m}(\mathbb{F})$. If $\text{rank}(I - A\Delta) < m$, then there exists $\mathbf{x} \in \mathbb{F}^m$ with $\mathbf{x} \neq \mathbf{0}$ such that $A\Delta\mathbf{x} = \mathbf{x}$. Thus,

$$\|\mathbf{x}\|_2 = \|A\Delta\mathbf{x}\|_2 \leq \|A\|_2 \|\Delta\mathbf{x}\|_2 = \sigma_1 \|\Delta\mathbf{x}\|_2,$$

which implies

$$\sigma_1^{-1} \leq \frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \|\Delta\|_2.$$

Since $\|\Delta^*\|_2 = \|\sigma_1^{-1} \mathbf{v}_1 \mathbf{u}_1^H\|_2 = \sigma_1^{-1}$, it suffices to show that $\text{rank}(I - A\Delta^*) < m$. However, this follows immediately from the fact that $(I - A\Delta^*)\mathbf{u}_1 = \mathbf{0}$; see Exercise 4.40. $\square$

As a corollary, we immediately get the *small gain theorem*, which is important in control theory.

Corollary 4.6.9 (Small Gain Theorem). *If $A \in M_n(\mathbb{F})$, then $I - A\Delta$ is non-singular, provided that $\|A\|_2\|\Delta\|_2 < 1$.*

Proof. The proof is Exercise 4.41. $\square$

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 4.1. A matrix $A \in M_n(\mathbb{F})$ is *nilpotent* if $A^k = 0$ for some $k \in \mathbb{N}$. Show that if λ is an eigenvalue of a nilpotent matrix, then $\lambda = 0$. Hint: Show that if λ is an eigenvalue of A , then λ^k is an eigenvalue of A^k .
- 4.2. Let $V = \text{span}(\{1, x, x^2\})$ be a subspace of the inner product space $L^2([0, 1]; \mathbb{R})$. Let D be the derivative operator $D : V \rightarrow V$ given by $D[p](x) = p'(x)$. Find all the eigenvalues and eigenspaces of D . What are their algebraic and geometric multiplicities?
- 4.3. Show that the characteristic polynomial of any 2×2 matrix has the form

$$p(\lambda) = \lambda^2 - \text{tr}(A)\lambda + \det(A).$$

- 4.4. Recall that a matrix $A \in M_n(\mathbb{F})$ is Hermitian if $A^H = A$ and skew-Hermitian if $A^H = -A$. Using Exercise 4.3, prove that
 - (i) an Hermitian 2×2 matrix has only real eigenvalues;
 - (ii) a skew-Hermitian 2×2 matrix has only imaginary eigenvalues.
- 4.5. Let $A, B \in M_{m \times n}(\mathbb{F})$ and $C \in M_n(\mathbb{F})$, and assume that λ is *not* an eigenvalue of C . Prove that λ is an eigenvalue of $C - B^H A$ if and only if $A(C - \lambda I)^{-1} B^H$ has an eigenvalue equal to 1. Why is it important that λ is not an eigenvalue

of C ? Hint: If $\mathbf{x} \in \mathbb{F}^n$ is an eigenvector of $C - B^H A$, consider the vector $\mathbf{y} = A\mathbf{x} \in \mathbb{F}^m$.

4.6. Prove Proposition 4.1.22.

4.7. Prove Proposition 4.1.23. Hint: Use Exercise 2.47.

4.8. Let V be the span of the set $S = \{\sin(x), \cos(x), \sin(2x), \cos(2x)\}$ in the vector space $C^\infty(\mathbb{R}; \mathbb{R})$.

(i) Prove that S is a basis for V .

(ii) Let D be the derivative operator. Write the matrix representation of D in the basis S .

(iii) Find two complementary D -invariant subspaces in V .

4.9. Prove that the right shift operator on ℓ^∞ has no one-dimensional invariant subspace (see Remark 4.2.8).

4.10. Assume that V is a vector space and T is a linear operator on V . Prove that if W is a T -invariant subspace of V , then the map $T' : V/W \rightarrow V/W$ given by $T'(\mathbf{v} + W) = T(\mathbf{v}) + W$ is a well-defined linear transformation.

4.11. Let W_1 and W_2 be complementary subspaces of the vector space V . A *reflection through W_1 along W_2* is a linear operator $R : V \rightarrow V$ such that $R(\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{w}_1 - \mathbf{w}_2$ for $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$. Prove that the following are equivalent:

(i) There exist complementary subspaces W_1 and W_2 of V such that R is a reflection through W_1 along W_2 .

(ii) R is an *involution*, that is, $R^2 = I$.

(iii) $V = \mathcal{N}(R - I) \oplus \mathcal{N}(R + I)$. Hint: We have that

$$\mathbf{v} = \frac{1}{2}(I - R)\mathbf{v} + \frac{1}{2}(I + R)\mathbf{v}.$$

4.12. Let L be the linear operator on $\mathbb{R}^2$ that reflects around the line $y = 3x$.

(i) Find two complementary L -invariant subspaces of V .

(ii) Choose a basis T for $\mathbb{R}^2$ consisting of one vector from each of the two complementary L -invariant subspaces and write the matrix representation of L in that basis.

(iii) Write the transition matrix C_{ST} from T to the standard basis S .

(iv) Write the matrix representation of L in the standard basis.

4.13. Let

$$A = \begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix}.$$

Compute the transition matrix P such that $P^{-1}AP$ is diagonal.

4.14. Prove that the linear transformation D of Exercise 4.2 is not semisimple.

4.15. Prove Theorem 4.3.12.

4.16. Let A be the matrix in Exercise 4.13 above.

- (i) Compute $\lim_{n \rightarrow \infty} A^n$ with respect to the 1-norm; that is, find a matrix B such that for any $\varepsilon > 0$ there exists an $N > 0$ with $\|A^k - B\|_1 < \varepsilon$ whenever $k > N$. Hint: Use Proposition 4.3.10.
- (ii) Repeat part (i) for the ∞ -norm and the Frobenius norm. Does the answer depend on the choice of norm? We discuss this further in Section 5.8.
- (iii) Find all the eigenvalues of the matrix $3I + 5A + A^3$. Hint: Consider using Theorem 4.3.12.

4.17. If $p(z)$ is the characteristic polynomial of a semisimple matrix $A \in M_n(\mathbb{F})$, then prove that $p(A) = 0$. We show in Chapter 12 that this theorem holds even if the matrix is not semisimple.

4.18. Prove: If λ is an eigenvalue of the $A \in M_n(\mathbb{F})$, then there exists a nonzero row vector $\mathbf{x}^T$ such that $\mathbf{x}^T A = \lambda \mathbf{x}^T$.

4.19. Let $A \in M_n(\mathbb{F})$ be a semisimple matrix.

- (i) Let ℓ^T be a left eigenvector of A (a row vector) with eigenvalue λ , and let $\mathbf{r}$ be a right eigenvector of A with eigenvalue ρ . Prove that $\ell^T \mathbf{r} = 0$ if $\lambda \neq \rho$.
- (ii) Prove that for any eigenvalue λ of A , there exist corresponding left and right eigenvectors ℓ^T and $\mathbf{r}$, respectively, associated with λ such that $\ell^T \mathbf{r} = 1$.
- (iii) Provide an example of a semisimple matrix A showing that even if $\lambda = \rho \neq 0$, there can exist a left eigenvector ℓ^T associated with λ and a right eigenvector $\mathbf{r}$ associated with ρ such that $\ell^T \mathbf{r} = 0$.

4.20. Prove Lemma 4.4.2.

4.21. Let $(V, \langle \cdot, \cdot \rangle)$ be a finite-dimensional inner product space, $[\mathbf{v}_1, \dots, \mathbf{v}_n] \subset V$ an orthonormal basis, and T an orthonormal operator on V . Prove that there is an orthonormal operator Q such that $Q\mathbf{v}_1, \dots, Q\mathbf{v}_n$ is an orthonormal eigenbasis of T . (This shows that there is an orthonormal change of basis Q which takes the original orthonormal basis to an orthonormal eigenbasis of T .) Hint: By the spectral theorem, there exists an orthonormal eigenbasis of T .

4.22. Let S and T be operators on a finite-dimensional inner product space $(V, \langle \cdot, \cdot \rangle)$ which commute, that is, $ST = TS$.

- (i) Show that every eigenspace of S is T -invariant; that is, for each eigenvalue λ of S , the λ -eigenspace $\Sigma_\lambda(S)$ satisfies $T\Sigma_\lambda(S) \subset \Sigma_\lambda(S)$.
- (ii) Show that if S is simple, then T is semisimple.

4.23. Let T be any invertible operator on a finite-dimensional inner product space $(W, \langle \cdot, \cdot \rangle)$.

- (i) Show that TT^* is self-adjoint and $\langle \mathbf{w}, TT^* \mathbf{w} \rangle \geq 0$ for all $\mathbf{w} \in W$.
- (ii) Show that TT^* has a self-adjoint square root S ; that is, $TT^* = S^2$ and $S^* = S$.

- (iii) Show that $S^{-1}T$ is orthonormal.
- (iv) Show that there exist orthonormal operators U and V such that U^*TV is diagonal.

4.24. Given $A \in M_n(\mathbb{C})$, define the *Rayleigh quotient* as

$$\rho(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2},$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{F}^n$. Show that the Rayleigh quotient can only take on real values for Hermitian matrices and only imaginary values for skew-Hermitian matrices.

4.25. Let $A \in M_n(\mathbb{C})$ be a normal matrix with eigenvalues $(\lambda_1, \dots, \lambda_n)$ and corresponding orthonormal eigenvectors $[\mathbf{x}_1, \dots, \mathbf{x}_n]$.

(i) Show that the identity matrix can be written $I = \mathbf{x}_1\mathbf{x}_1^H + \dots + \mathbf{x}_n\mathbf{x}_n^H$.
Hint: What is $(\mathbf{x}_1\mathbf{x}_1^H + \dots + \mathbf{x}_n\mathbf{x}_n^H)\mathbf{x}_j$?

(ii) Show that A can be written as $A = \lambda_1\mathbf{x}_1\mathbf{x}_1^H + \dots + \lambda_n\mathbf{x}_n\mathbf{x}_n^H$. This is called an *outer product expansion*.

4.26.† Let $A, B \in M_n(\mathbb{F})$ be Hermitian, and let $\mathbb{F}^n$ be equipped with the standard inner product. If $[\mathbf{x}_1, \dots, \mathbf{x}_n]$ is an orthonormal eigenbasis of B , then prove that

$$\operatorname{tr}(AB) = \sum_{i=1}^n \langle \mathbf{x}_i, AB\mathbf{x}_i \rangle.$$

Hint: Use Exercises 2.25 and 4.25.

4.27. Assume $A \in M_n(\mathbb{F})$ is positive definite. Prove that all its diagonal entries are real and positive.

4.28. Assume $A, B \in M_n(\mathbb{F})$ are positive semidefinite. Prove that

$$0 \leq \operatorname{tr}(AB) \leq \operatorname{tr}(A)\operatorname{tr}(B),$$

and use this result to prove that $\|\cdot\|_F$ is a matrix norm.

4.29. Let

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 \end{bmatrix}.$$

- (i) Find the eigenvalues of $A^H A$ along with their algebraic and geometric multiplicities.
 - (ii) Find the singular values of A .
 - (iii) Compute the entire SVD of A .
 - (iv) Give an orthonormal basis for each of the four fundamental subspaces of A .
- 4.30. Let $V = \operatorname{span}(\{1, x, x^2\})$ be a subspace of the inner product space $L^2([0, 1]; \mathbb{R})$. Let D be the derivative operator $D: V \rightarrow V$ given by $D[p](x) = p'(x)$. Write the matrix representation of D with respect to the basis $[1, x, x^2]$ and compute its SVD.

- 4.31.  Assume $A \in M_{m \times n}(\mathbb{F})$ and A is not identically zero. Prove that
- (i) $\|A\|_2 = \sigma_1$, where σ_1 is the largest singular value of A ;
 - (ii) if A is invertible, then $\|A^{-1}\|_2 = \sigma_n^{-1}$;
 - (iii) $\|A^H\|_2^2 = \|A^\top\|_2^2 = \|A^H A\|_2 = \|A\|_2^2$;
 - (iv) if $U \in M_m(\mathbb{F})$ and $V \in M_n(\mathbb{F})$ are orthonormal, then $\|UAV\|_2 = \|A\|_2$.
- 4.32.  Assume $A \in M_{m \times n}(\mathbb{F})$ is of rank r . Prove that
- (i) if $U \in M_m(\mathbb{F})$ and $V \in M_n(\mathbb{F})$ are orthonormal, then $\|UAV\|_F = \|A\|_F$;
 - (ii) $\|A\|_F = (\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2)^{1/2}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of A .
- 4.33. Assume $A \in M_n(\mathbb{F})$. Prove that

$$\|A\|_2 = \sup_{\substack{\|x\|_2=1 \\ \|y\|_2=1}} |y^H A x|. \quad (4.26)$$

Hint: Use Exercise 4.31.

- 4.34.* Let B and D be Hermitian. Prove that the matrix

$$A = \begin{bmatrix} B & C \\ C^H & D \end{bmatrix}$$

is positive definite if and only if $B > 0$ and $D - C^H B^{-1} C > 0$. Hint: Find a matrix P of the form $\begin{bmatrix} I & E \\ 0 & I \end{bmatrix}$ that makes $P^H A P$ block diagonal.

- 4.35. Let $A \in M_n(\mathbb{F})$ be nonsingular. Prove that the modulus of the determinant is the product of the singular values:

$$|\det A| = \prod_{i=1}^n \sigma_i.$$

- 4.36. Give an example of a 2×2 matrix whose determinant is nonzero and whose singular values are not equal to any of its eigenvalues.
- 4.37. Let A be the matrix in Exercise 4.29. Find the Moore–Penrose inverse $A^\dagger$ of A . Compare $A^\dagger A$ to $A^H A$.
- 4.38. Prove Proposition 4.6.2.
- 4.39. Prove Corollary 4.6.7.
- 4.40. Finish the proof of Theorem 4.6.8 by showing that when $A \in M_{m \times n}(\mathbb{F})$ has SVD equal to $A = U \Sigma V^H$, and if $\Delta^* = \sigma_1^{-1} \mathbf{v}_1 \mathbf{u}_1^H$, then $(I - A \Delta^*) \mathbf{u}_1 = \mathbf{0}$.
- 4.41.* Prove the small gain theorem (Corollary 4.6.9). Hint: What is the contrapositive of Theorem 4.6.8?
- 4.42.* Let $A, \Delta \in M_n(\mathbb{F})$ with $\text{rank}(A) = r$.
- (i) If $\mathcal{R}(\Delta) \subset \mathcal{R}(A)$, prove that $A + \Delta = A(I + A^\dagger \Delta)$.
 - (ii) For any $X, Y \in M_n(\mathbb{F})$, prove that if $\text{rank}(XY) < \text{rank}(X)$, then $\text{rank}(Y) < n$.

- (iii) Use the first parts of this exercise and Theorem 4.6.8 to prove that if $A + \Delta$ has rank $s < r$, then $\|\Delta\|_2 \geq \sigma_r$, without using the Schmidt, Mirsky, Eckart–Young theorems.
-

Notes

We have focused primarily on spectral theory of finite-dimensional operators because the infinite-dimensional case is very different and has many subtleties. This is normally covered in a functional analysis course. Some resources for the infinite-dimensional case include [Pro08, Con90, Rud91].

Part II

Nonlinear Analysis I

5

Metric Space Topology

The angel of topology and the devil of abstract algebra fight for the soul of each individual mathematical domain.

—Herman Weyl

In single-variable analysis, the distance between any two numbers $x, y \in \mathbb{F}$ is defined to be $|x - y|$, where $|\cdot|$ is the absolute value in $\mathbb{R}$ or the modulus in $\mathbb{C}$. The distance function allows us to rigorously define several essential ideas in mathematical analysis, such as continuity and convergence. In Chapter 3, we generalized the distance function to normed linear spaces, taking the distance between two vectors to be the norm of their difference. While normed linear spaces are central to applied mathematics, there are many important problems that do not have the algebraic or geometric structure needed to define a norm. In this chapter, we consider a more general notion of distance that allows us to extend some of the most important ideas of mathematical analysis to more abstract settings.

We begin by defining a distance function, or *metric*, on an abstract space X as a nonnegative, symmetric, real-valued function on $X \times X$ that satisfies the triangle inequality and a certain positivity condition. This gives us the necessary framework to quantify “nearness” and define continuity, convergence, and many other important concepts. The metric also allows us to generalize the notions of open and closed intervals from single-variable real analysis to open and closed sets in the space X .

The collection of all open sets in a given space is called the *topology* of that space, and any properties of a space that depend only on the open sets are called *topological properties*. Two of the most important topological properties are *compactness* and *connectedness*. Compactness implies that every sequence has a convergent subsequence, and connectedness means that the set cannot be broken apart into two or more separate pieces. Continuous functions have the very nice property that they preserve both compactness and connectedness.

Cauchy sequences, which have the property that the terms of the sequence get arbitrarily close to each other as the sequence progresses, are some of the most important sequences in a metric space. If every Cauchy sequence converges in the space, we say that the space is *complete*. Roughly speaking, this means any

point that we can get arbitrarily close to must actually be in the space—there are no holes or gaps in the space. For example, $\mathbb{Q}$ is not complete because we can approximate irrational numbers (not in $\mathbb{Q}$) as closely as we like with rational numbers. Many of the most useful spaces in mathematical analysis are complete; for example, $(\mathbb{F}^n, \|\cdot\|_p)$ is complete for any $n \in \mathbb{N}$ and any $p \in [1, \infty]$.

When normed vector spaces are also complete, they are called *Banach spaces*. Banach spaces are very important in analysis and applied mathematics, and most of the normed linear spaces used in applied mathematics are Banach spaces. Some important examples of Banach spaces include $\mathbb{F}^n$, the space of matrices $M_{m \times n}(\mathbb{F})$ (which is isomorphic to $\mathbb{F}^{nm}$), the space of continuous functions $(C([a, b]; \mathbb{R}), \|\cdot\|_{L^\infty})$, the space of bounded functions, and the spaces ℓ^p for $1 \leq p \leq \infty$.

Although this chapter is a little more abstract than the rest of the book so far, and even though we do not give as many immediate applications of the ideas in this chapter, the material in this chapter is fundamental to applied mathematics and provides powerful tools that you can use repeatedly throughout the rest of the book and beyond.

5.1 Metric Spaces and Continuous Functions

In this section, we define a notion of distance, called a *metric*, on a set. Sets that have a metric are called *metric spaces*.³⁴ The metric allows us to define open sets and define what it means for a function to be continuous.

5.1.1 Metric Spaces

Definition 5.1.1. A metric on a set X is a map $d : X \times X \rightarrow \mathbb{R}$ that satisfies the following properties for all $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ in X :

- (i) *Positive definiteness:* $d(\mathbf{x}, \mathbf{y}) \geq 0$, with $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
- (ii) *Symmetry:* $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
- (iii) *Triangle Inequality:* $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

The pair (X, d) is called a *metric space*.

Example 5.1.2. Perhaps the most common metric space is the *Euclidean* metric on $\mathbb{F}^n$ given by the 2-norm, that is, $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$. Unless we specifically say otherwise, we always use this metric on $\mathbb{F}^n$.

Example 5.1.3. The reader should check that each of the examples below satisfies the definition of a metric space.

³⁴Metric spaces are not necessarily vector spaces. Indeed, there need not be any binary operations defined on a metric space.

- (i) We can generalize Example 5.1.2 to more general normed linear spaces. By Definition 3.1.11, any norm $\|\cdot\|$ on a vector space induces a natural metric

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|. \quad (5.1)$$

Unless we specifically say otherwise, we always use this metric on a normed linear space.

- (ii) For $f, g \in C([a, b]; \mathbb{R})$ and for any $p \in [1, \infty]$, we have the metric

$$d^p(f, g) = \begin{cases} \left(\int_a^b |f(t) - g(t)|^p dt \right)^{1/p}, & 1 \leq p < \infty, \\ \sup_{x \in [a, b]} |f(t) - g(t)|, & p = \infty. \end{cases} \quad (5.2)$$

These are also written as $\|f - g\|_{L^p}$ and $\|f - g\|_{L^\infty}$, respectively.

- (iii) The *discrete metric* on X is

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } \mathbf{x} = \mathbf{y}, \\ 1 & \text{if } \mathbf{x} \neq \mathbf{y}. \end{cases} \quad (5.3)$$

Thus, no two distinct points are close together—they are always the same distance apart.

- (iv) Let $((X_i, d_i))_{i=1}^n$ be a collection of metric spaces, and let $X = X_1 \times X_2 \times \cdots \times X_n$ be the Cartesian product. For any two points $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ in X , define

$$d^p(\mathbf{x}, \mathbf{y}) = \begin{cases} \left(\sum_{i=1}^n d_i(x_i, y_i)^p \right)^{1/p}, & 1 \leq p < \infty, \\ \sup_i d_i(x_i, y_i), & p = \infty. \end{cases} \quad (5.4)$$

This defines a metric on X called the p -metric. If the metric d_i on each X_i is induced from a norm $\|\cdot\|_{X_i}$, as in (i) above, then the p -metric on X is the metric induced by the p -norm (see Exercise 3.34).

Example 5.1.4. Let (X, d) be a metric space. We can create a new metric on X :

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{d(\mathbf{x}, \mathbf{y})}{1 + d(\mathbf{x}, \mathbf{y})}, \quad (5.5)$$

where no two points are farther apart than 1. To show that (5.5) is a metric on X , see Exercise 5.4.

Remark 5.1.5. Every norm induces a metric space, but not every metric space is a normed space. For example, a metric can be defined on sets that are not vector spaces. And even if the underlying space is a vector space, we can install metrics on it that are not induced by a norm.

5.1.2 Open Sets

Those familiar with single-variable analysis know that open intervals in $\mathbb{R}$ are important for defining continuous functions, differentiability, and local properties of functions. Open sets in metric spaces allow us to define similar ideas in an abstract metric space.

Throughout the remainder of this section, let (X, d) be a metric space.

Definition 5.1.6. For each point $\mathbf{x}_0 \in X$ and $r > 0$, define the open ball with center at $\mathbf{x}_0$ and radius $r > 0$ to be the set

$$B(\mathbf{x}_0, r) = \{\mathbf{x} \in X \mid d(\mathbf{x}, \mathbf{x}_0) < r\}.$$

Definition 5.1.7. A subset $E \subset X$ is a neighborhood of a point $\mathbf{x} \in X$ if there exists an open ball $B(\mathbf{x}, r) \subset E$. In this case we say that $\mathbf{x}$ is an interior point of E . We write E° to denote the set of interior points of E .

Definition 5.1.8. A subset $E \subset X$ is an open set if every point $\mathbf{x} \in E$ is an interior point of E .

Example 5.1.9. Both X and $\emptyset$ are open sets. First, X is open since $B(\mathbf{x}, r) \subset X$ for all $\mathbf{x} \in X$ and for all $r > 0$. That $\emptyset$ is open follows vacuously—every point in $\emptyset$ satisfies the condition because there are no points in $\emptyset$.

Example 5.1.10. In the Euclidean norm, $B(\mathbf{x}_0, r)$ looks like a ball, which is why we call it the “open ball.” However, in other metrics open balls can take on very different shapes. For example, in $\mathbb{R}^3$ the open ball in the metric induced by the 1-norm is an octahedron, and the open ball in the metric induced by the ∞ -norm is a cube (see Figure 3.6). In the discrete metric (see Example 5.1.3(iii)), open balls of radius one or less are just points (singleton sets); that is, $B(\mathbf{x}, 1) = \{\mathbf{x}\}$ for each $\mathbf{x} \in X$, while $B(\mathbf{x}, r) = X$ for all $r > 1$.

Example 5.1.11. Another important example is the space $C([0, 1]; \mathbb{F})$ with the metric $d(f, g) = \|f - g\|_{L^\infty}$ determined by the sup norm $\|\cdot\|_{L^\infty}$. In this space the open ball $B(0, 1)$ around the zero function is infinite dimensional,

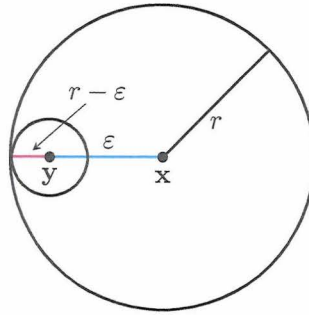


Figure 5.1. A ball inside of a ball, as discussed in the proof of Theorem 5.1.12. Here $d(\mathbf{y}, \mathbf{x}) = \varepsilon$, and the distance from $\mathbf{y}$ to the edge of the ball is $r - \varepsilon$.

so it is hard to draw, but it consists of all functions that only take on values with modulus less than 1 on the interval $[0, 1]$. It contains $\sin(x)$ because

$$\|\sin(x) - 0\|_{L^\infty} = \sup_{x \in [0,1]} |\sin(x)| = \sin(1) < \sin(\pi/3) < 1.$$

But it does not contain $\cos(x)$ because

$$\|\cos(x) - 0\|_{L^\infty} = \sup_{x \in [0,1]} |\cos(x)| = 1.$$

We now prove that the balls defined in Definition 5.1.6 are, in fact, open sets, as defined in Definition 5.1.8.

Theorem 5.1.12. *If $\mathbf{y} \in B(\mathbf{x}, r)$ for some $\mathbf{x} \in X$, then $B(\mathbf{y}, r - \varepsilon) \subset B(\mathbf{x}, r)$, where $\varepsilon = d(\mathbf{x}, \mathbf{y})$.*

Proof. Assume $\mathbf{z} \in B(\mathbf{y}, r - \varepsilon)$. Thus, $d(\mathbf{z}, \mathbf{y}) < r - \varepsilon$, so $d(\mathbf{z}, \mathbf{y}) + \varepsilon < r$. This implies that $d(\mathbf{z}, \mathbf{y}) + d(\mathbf{y}, \mathbf{x}) < r$, which by the triangle inequality yields $d(\mathbf{z}, \mathbf{x}) < r$, or equivalently $\mathbf{z} \in B(\mathbf{x}, r)$ (see Figure 5.1). $\square$

Example 5.1.13.* Consider the metric space $(\mathbb{F}^n, d)$, where d is the Euclidean metric. Let $\mathbf{e}_i$ be the i th standard basis vector. Note that $d(\mathbf{e}_i, \mathbf{e}_j) = \sqrt{2}$ whenever $i \neq j$. We can show that each basis element $\mathbf{e}_i$ is contained in exactly one ball in $\mathcal{C} = \{B(\mathbf{e}_i, \sqrt{2})\}_{i=1}^n \cup \{B(-\mathbf{e}_i, \sqrt{2})\}_{i=1}^n$ and that each element $\mathbf{x}$ in the unit ball $B(0, 1)$ is contained in at least one of the $2n$ balls in $\mathcal{C}$ (see Figure 5.2 and Exercise 5.7).

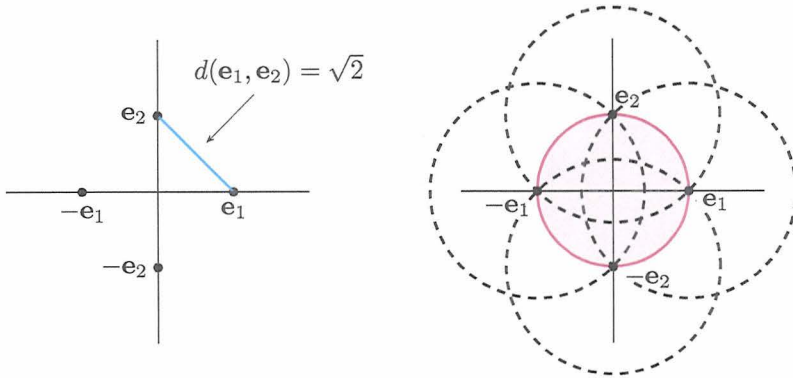


Figure 5.2. Illustration of Example 5.1.13 in $\mathbb{R}^2$. Each basis element e_i is contained in exactly one ball in $\mathcal{C} = \{B(e_i, \sqrt{2})\}_{i=1}^2 \cup \{B(-e_i, \sqrt{2})\}_{i=1}^2$, and each point in the (red) unit ball $B(0, 1)$ is contained in at least one of the 4 balls (dashed) in $\mathcal{C}$.

By the definition of an open set, any open set can be written as a union of open balls. We now show that any union of open sets is open and that the intersection of a *finite* number of open sets is open.

Theorem 5.1.14. *The union of any collection of open sets is open, and the intersection of any finite collection of open sets is open.*

Proof. We first prove the result for unions. Let $(G_\alpha)_{\alpha \in J}$ be a collection of open sets G_α indexed by the set J . If $\mathbf{x} \in \bigcup_{\alpha \in J} G_\alpha$, then $\mathbf{x} \in G_\alpha$ for some $\alpha \in J$. Hence, there exists $\varepsilon > 0$ such that $B(\mathbf{x}, \varepsilon) \subset G_\alpha$, and so $B(\mathbf{x}, \varepsilon) \subset \bigcup_{\alpha \in J} G_\alpha$. Therefore $\bigcup_{\alpha \in J} G_\alpha$ is open.

We now prove the result for intersections. Let $(G_k)_{k=1}^n$ be a finite collection of open sets. If $\mathbf{x} \in \bigcap_{k=1}^n G_k$, then for each k there exists $\varepsilon_k > 0$ such that $B(\mathbf{x}, \varepsilon_k) \subset G_k$. Let $\varepsilon = \min\{\varepsilon_1, \dots, \varepsilon_n\}$, which is positive. Thus, $B(\mathbf{x}, \varepsilon) \subset G_k$ for each k , and so $B(\mathbf{x}, \varepsilon) \subset \bigcap_{k=1}^n G_k$. It follows that $\bigcap_{k=1}^n G_k$ is open. $\square$

Example 5.1.15. Note that an infinite intersection of open sets need not be open. As a simple example, consider the following intersection of open sets in $\mathbb{R}$ with the usual metric:

$$\bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n} \right).$$

The intersection is just the single point $\{0\}$, which is not an open set in $\mathbb{R}$.

Theorem 5.1.16. *The following properties hold for any subset E of X :*

- (i) $(E^\circ)^\circ = E^\circ$, and hence E° is open.

- (ii) If G is an open subset of E , then $G \subset E^\circ$.
- (iii) E is open if and only if $E = E^\circ$.
- (iv) E° is the union of all open sets contained in E .

Proof.

- (i) By definition $(E^\circ)^\circ \subset E^\circ$. Conversely, if $\mathbf{x} \in E^\circ$, then there exists an open ball $B(\mathbf{x}, \delta) \subset E$. By Theorem 5.1.12 every point $\mathbf{y} \in B(\mathbf{x}, \delta)$ is contained in $B(\mathbf{x}, \delta)^\circ \subset E$. Hence, $B(\mathbf{x}, \delta) \subset E^\circ$, which implies $\mathbf{x} \in (E^\circ)^\circ$.
- (ii) Assume G is an open subset of E . If $\mathbf{x} \in G$, then there exists $\varepsilon > 0$ such that $B(\mathbf{x}, \varepsilon) \subset G$, which implies that $B(\mathbf{x}, \varepsilon) \subset E$; it follows that $\mathbf{x} \in E^\circ$. Thus, $G \subset E^\circ$.
- (iii) See Exercise 5.6.
- (iv) Let $(G_\alpha)_{\alpha \in J}$ be the collection of all open sets contained in E . By (ii), we have that $G_\alpha \subset E^\circ$ for all $\alpha \in J$. Thus, $\bigcup_{\alpha \in J} G_\alpha \subset E^\circ$. On the other hand, if $\mathbf{x} \in E^\circ$, then by definition it is contained in an open ball $B(\mathbf{x}, \varepsilon) \subset E$, which by Theorem 5.1.12 is an open subset of E . $\square$

5.2 Continuous Functions and Limits

The concept of a continuous function is fundamental in mathematics. There are many different ways to define continuous functions on metric spaces. We give one of these definitions and then show that it agrees with some of the others, including the familiar definition in terms of limits (see Theorem 5.2.9).

Throughout this section let (X, d) and (Y, ρ) be metric spaces.

5.2.1 Continuous Functions

Definition 5.2.1. A function $f : X \rightarrow Y$ is continuous at a point $\mathbf{x}_0 \in X$ if for all $\varepsilon > 0$ there exists $\delta > 0$ such that $\rho(f(\mathbf{x}), f(\mathbf{x}_0)) < \varepsilon$ whenever $d(\mathbf{x}, \mathbf{x}_0) < \delta$. A function $f : X \rightarrow Y$ is continuous on a subset $E \subset X$ if it is continuous at each $\mathbf{x}_0 \in E$. The set of continuous functions from X to Y is denoted $C(X; Y)$.

Example 5.2.2. We have the following examples of continuous functions:

- (i) Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = |x - y|$. We show that f is continuous at $(0, 0)$. Note that

$$|f(x, y) - f(0, 0)| = |x - y| \leq |x| + |y| \leq 2\|(x, y)\|_2 = 2d((x, y), \mathbf{0}).$$

Setting $\delta = \varepsilon/2$, we have $|f(x, y) - f(0, 0)| < \varepsilon$ whenever $\|(x, y)\|_2 < \delta$.

- (ii)  If $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ are normed linear spaces, then every bounded^a linear transformation $T \in \mathcal{B}(V, W)$ is continuous at each $\mathbf{x}_0 \in V$. Given $\varepsilon > 0$, we set $\delta = \varepsilon/(1 + \|T\|_{V,W})$. Hence,

$$\|T(\mathbf{x}) - T(\mathbf{x}_0)\|_W = \|T(\mathbf{x} - \mathbf{x}_0)\|_W \leq \|T\|_{V,W} \|\mathbf{x} - \mathbf{x}_0\|_V < \varepsilon$$

whenever $\|\mathbf{x} - \mathbf{x}_0\|_V < \delta$. In other words, $\mathcal{B}(V; W) \subset C(V; W)$. On the other hand, an unbounded linear transformation is nowhere continuous. In other words, it is not continuous at any point on its domain (see Exercise 5.21).

- (iii)  A function $f : X \rightarrow Y$ is *Lipschitz continuous* (or just *Lipschitz* for short) if there exists $K > 0$ such that $\rho(f(\mathbf{x}_1), f(\mathbf{x}_2)) \leq Kd(\mathbf{x}_1, \mathbf{x}_2)$ for all $\mathbf{x}_1, \mathbf{x}_2 \in X$. Every Lipschitz continuous function is continuous on all of X (set $\delta = \varepsilon/K$).
- (iv)  For fixed $\mathbf{x}_0 \in X$, the map $f : X \rightarrow \mathbb{R}$ given by $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{x}_0)$ is continuous at each $\mathbf{x} \in X$. To see this, for any $\varepsilon > 0$ set $\delta = \varepsilon$, and thus

$$|f(\mathbf{x}) - f(\mathbf{y})| = |d(\mathbf{x}, \mathbf{x}_0) - d(\mathbf{y}, \mathbf{x}_0)| \leq d(\mathbf{x}, \mathbf{y}) < \varepsilon,$$

whenever $d(\mathbf{x}, \mathbf{y}) < \delta$. Note that this is similar to the solution of Exercise 3.16.

- (v) Consider the space $(\mathbb{F}^n, d)$, where d is the Euclidean metric $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$. For each k the k th projection map $\pi_k : \mathbb{F}^n \rightarrow \mathbb{F}$, defined in Example 2.1.3(i), is continuous at each $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{F}^n$. Given any ε , setting $\delta = \varepsilon$ gives $|\pi_k(\mathbf{x}) - \pi_k(\mathbf{y})| = |x_k - y_k| \leq d(\mathbf{x}, \mathbf{y}) < \varepsilon$, whenever $d(\mathbf{x}, \mathbf{y}) < \delta$. Note that this is a bounded linear transformation and thus is a special case of (ii).

^aSee Definition 3.5.10.

5.2.2 Alternative Definitions of Continuity

The next theorem is a very important result, giving an alternative definition of continuity on a set in terms of the function (or rather its preimage) preserving open sets. This tells us essentially that continuous functions are to the study of open sets (the subject called *topology*) what linear transformations are to linear algebra. A general philosophy in modern mathematics is that a collection of sets with some special structure (like vector spaces, or metric spaces) is best understood by studying the functions that preserve the special structure. Continuous functions are those functions for spaces with open sets (called *topological spaces*).

Theorem 5.2.3. *A function $f : X \rightarrow Y$ is continuous on X if and only if the preimage $f^{-1}(U)$ of every open set $U \subset Y$ is open in X .*

Nota Bene 5.2.4. Recall that the notation $f^{-1}(U)$ does not mean that f has an inverse. The set $f^{-1}(U) = \{\mathbf{x} \in X \mid f(\mathbf{x}) \in U\}$ always exists (but may be empty), even if f has no inverse.

Proof. ($\Rightarrow$) Assume that f is continuous on X and $U \subset Y$ is open. For $\mathbf{x}_0 \in f^{-1}(U)$, choose $\varepsilon > 0$ so that $B(f(\mathbf{x}_0), \varepsilon) \subset U$. Since f is continuous at $\mathbf{x}_0$, there exists $\delta > 0$ such that $f(\mathbf{x}) \in B(f(\mathbf{x}_0), \varepsilon)$ whenever $\mathbf{x} \in B(\mathbf{x}_0, \delta)$, or, in other words, $f(B(\mathbf{x}_0, \delta)) \subset B(f(\mathbf{x}_0), \varepsilon)$. Hence, $B(\mathbf{x}_0, \delta) \subset f^{-1}(U)$. Since this holds for all $\mathbf{x}_0 \in f^{-1}(U)$, we have that $f^{-1}(U)$ is open.

($\Leftarrow$) Let $\varepsilon > 0$ be given. For any $\mathbf{x}_0 \in X$, the ball $B(f(\mathbf{x}_0), \varepsilon)$ is open, and by hypothesis, so is its preimage $f^{-1}(B(f(\mathbf{x}_0), \varepsilon))$. Thus, there exists $\delta > 0$ such that $B(\mathbf{x}_0, \delta) \subset f^{-1}(B(f(\mathbf{x}_0), \varepsilon))$, and hence f is continuous at $\mathbf{x}_0 \in X$. $\square$

The following is an easy consequence of Theorem 5.2.3.

Corollary 5.2.5. *Compositions of continuous maps are continuous; that is, if $f : (X, d) \rightarrow (Y, \rho)$ and $g : (Y, \rho) \rightarrow (Z, \tau)$ are continuous mappings on X and Y , respectively, then $h = g \circ f : (X, d) \rightarrow (Z, \tau)$ is continuous.*

Proof. If $U \subset Z$ is open, then $g^{-1}(U)$ is open in Y , which implies that $f^{-1}(g^{-1}(U))$ is open in X . Hence, $h^{-1}(U)$ is open whenever U is open. $\square$

Theorem 5.2.3 suggests yet another alternative definition of continuity at a point.

Proposition 5.2.6. *A function $f : X \rightarrow Y$ is continuous at $\mathbf{x}_0 \in X$ if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B(\mathbf{x}_0, \delta)) \subset B(f(\mathbf{x}_0), \varepsilon)$.*

Proof. This follows from the definition of continuity by just writing out the definition of the various open balls. More precisely, $\mathbf{x} \in B(\mathbf{x}_0, \delta)$ if and only if $d(\mathbf{x}, \mathbf{x}_0) < \delta$; and $f(\mathbf{x}) \in B(f(\mathbf{x}_0), \varepsilon)$ if and only if $\rho(f(\mathbf{x}), f(\mathbf{x}_0)) < \varepsilon$. $\square$

Proposition 5.2.7. *Consider the space $\mathbb{F}^n$ with the metric d_p induced by the p -norm for some fixed $p \in [1, \infty]$. Let $\|\cdot\|$ be any norm on $\mathbb{F}^n$. The function $f : \mathbb{F}^n \rightarrow \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{x}\|$ is continuous with respect to the metric d_p . In particular, it is continuous with respect to the usual metric ($p = 2$).*

Proof. Let $M = \max(\|\mathbf{e}_1\|, \dots, \|\mathbf{e}_n\|)$, where $\mathbf{e}_i$ is the i th standard basis vector. Let

$$q = \begin{cases} \frac{p}{p-1}, & p \in (1, \infty], \\ \infty, & p = 1. \end{cases}$$

Note that $\frac{1}{p} + \frac{1}{q} = 1$, and for any $\mathbf{z} = (z_1, \dots, z_n) = \sum_{i=1}^n z_i \mathbf{e}_i$, the triangle inequality together with Hölder's inequality (Corollary 3.6.4) gives

$$f(\mathbf{z}) = \|\mathbf{z}\| \leq \sum_{i=1}^n |z_i| \|\mathbf{e}_i\| \leq \left(\sum_{i=1}^n |z_i|^p \right)^{1/p} \left(\sum_{i=1}^n \|\mathbf{e}_i\|^q \right)^{1/q} \leq n^{1/q} M \|\mathbf{z}\|_p.$$

Therefore, for any $\varepsilon > 0$ we have

$$d(f(\mathbf{x}), f(\mathbf{y})) = \left| \|\mathbf{x}\| - \|\mathbf{y}\| \right| \leq \|\mathbf{x} - \mathbf{y}\| \leq n^{1/q} M \|\mathbf{x} - \mathbf{y}\|_p < \varepsilon,$$

whenever $d_p(\mathbf{x} - \mathbf{y}) < \varepsilon/(n^{1/q}M)$. This shows f is continuous with respect to d_p . $\square$

5.2.3 Limits of Functions

The ideas of the limit of a function and the limit of a sequence are extremely powerful. In this section we define a limit of a function on an arbitrary metric space.

Definition 5.2.8. Let $f : X \rightarrow Y$ be a function. The point $\mathbf{y}_0 \in Y$ at $\mathbf{x}_0 \in X$ is called the limit of f at $\mathbf{x}_0 \in X$ if for all $\varepsilon > 0$ there exists $\delta > 0$ such that $\rho(f(\mathbf{x}), \mathbf{y}_0) < \varepsilon$, whenever $0 < d(\mathbf{x}, \mathbf{x}_0) < \delta$. We denote the limit as $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x}) = \mathbf{y}_0$.

Theorem 5.2.9. A function $f : X \rightarrow Y$ is continuous at $\mathbf{x}_0 \in X$ if and only if $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$.

Proof. This is immediate from Definition 5.2.1. $\square$

Example 5.2.10. The function

$$f(x, y) = \begin{cases} \frac{x^2 y^3}{(x^2 + y^2)^2}, & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0), \end{cases}$$

is continuous at zero. To see this, note that $|x| \leq (x^2 + y^2)^{1/2}$ and $|y| \leq (x^2 + y^2)^{1/2}$, and thus $|x^2 y^3| \leq (x^2 + y^2)^{5/2}$. This gives $|f(x, y) - 0| \leq (x^2 + y^2)^{1/2}$, and thus $\lim_{(x, y) \rightarrow (0, 0)} f(x, y) = 0$.

We conclude this subsection by showing that sums and products of continuous functions are also continuous.

Proposition 5.2.11. If V is a vector space, and if $f : X \rightarrow V$ and $g : X \rightarrow V$ are continuous, then so is the sum $f + g : X \rightarrow V$. If $h : X \rightarrow \mathbb{F}$ is continuous, then the scalar product $hf : X \rightarrow V$ is also continuous.

Proof. We prove the case of the product hf . The proof of the sum is similar and is left to the reader.

Let $\varepsilon > 0$ be given. Since f is continuous at $\mathbf{x}_0$, there exists $\delta_1 > 0$ so that $\|f(\mathbf{x}) - f(\mathbf{x}_0)\| < \frac{\varepsilon}{2(|h(\mathbf{x}_0)|+1)}$ whenever $d(\mathbf{x}, \mathbf{x}_0) < \delta_1$. Since h is continuous, choose

$\delta_2 > 0$ so that $|h(\mathbf{x}) - h(\mathbf{x}_0)| < \min(1, \frac{\varepsilon}{2(\|f(\mathbf{x}_0)\|+1)})$ whenever $d(\mathbf{x}, \mathbf{x}_0) < \delta_2$. Note that $|h(\mathbf{x}) - h(\mathbf{x}_0)| < 1$ implies $|h(\mathbf{x})| < |h(\mathbf{x}_0)| + 1$, so we have

$$\begin{aligned} \|(hf)(\mathbf{x}) - (hf)(\mathbf{x}_0)\| &\leq \|h(\mathbf{x})f(\mathbf{x}) - h(\mathbf{x})f(\mathbf{x}_0) + h(\mathbf{x})f(\mathbf{x}_0) - h(\mathbf{x}_0)f(\mathbf{x}_0)\| \\ &\leq \|f(\mathbf{x}) - f(\mathbf{x}_0)\| |h(\mathbf{x})| + \|f(\mathbf{x}_0)\| |h(\mathbf{x}) - h(\mathbf{x}_0)| \\ &< \|f(\mathbf{x}) - f(\mathbf{x}_0)\| (|h(\mathbf{x}_0)| + 1) + \|f(\mathbf{x}_0)\| |h(\mathbf{x}) - h(\mathbf{x}_0)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

whenever $d(\mathbf{x}, \mathbf{x}_0) < \delta = \min\{\delta_1, \delta_2\}$. $\square$

5.2.4 Limit Points of Sets

We now define the important concept of a limit point of a set.

Definition 5.2.12. A point $\mathbf{p} \in X$ is a limit point of a set $E \subset X$ if every neighborhood of $\mathbf{p}$ intersects $E \setminus \{\mathbf{p}\}$.

Nota Bene 5.2.13. Despite the misleading name, limit points of a set are not the same as the limit of a function. In the following section we also define limits of a sequence, which is yet another concept with almost the same name. We would prefer to use very distinct names for these very distinct ideas, but in order to communicate with other analysts, you need to know them by the standard (and sometimes confusing) names.

Example 5.2.14. Any point on the disk $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$ is a limit point of the open ball $B(0, 1)$ in $\mathbb{R}^2$, with the Euclidean metric.

Definition 5.2.15. If $E \subset X$ and $\mathbf{p} \in E$ is not a limit point of E , we say that $\mathbf{p}$ is an isolated point of E .

Example 5.2.16.

- (i) For any subset E of a space with the discrete metric, each point is an isolated point of E .
- (ii) Each element of the set $\mathbb{Z} \times \mathbb{Z} \subset \mathbb{R} \times \mathbb{R}$ is an isolated point of $\mathbb{Z} \times \mathbb{Z}$. For any $\mathbf{p} = (m, n) \in \mathbb{Z} \times \mathbb{Z}$, it is easy to see that $B(\mathbf{p}, 1) \setminus \{\mathbf{p}\}$ does not intersect $\mathbb{Z} \times \mathbb{Z}$.

Definition 5.2.17. Let $E \subset X$. We say that E is dense in X if every point in X is either in E or is a limit point of E .

Example 5.2.18. The set $\mathbb{Q} \times \mathbb{Q}$ is dense in $\mathbb{R}^2$. It is easy to see that this generalizes: $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$.

Theorem 5.2.19. If $\mathbf{p}$ is a limit point of $E \subset X$, then every neighborhood of $\mathbf{p}$ contains infinitely many points of E .

Proof. Suppose there is a neighborhood N of $\mathbf{p}$ that contains only a finite number of elements of $E \setminus \{\mathbf{p}\}$, say $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. If $\varepsilon = \min_k d(\mathbf{p}, \mathbf{x}_k)$, then $B(\mathbf{p}, \varepsilon)$ does not intersect $E \setminus \{\mathbf{p}\}$. This contradicts the fact that $\mathbf{p}$ is a limit point of E . Thus, $N \cap (E \setminus \{\mathbf{p}\})$ contains infinitely many points. $\square$

Remark 5.2.20. An immediate consequence of the theorem is that a finite set has no limit points.

5.3 Closed Sets, Sequences, and Convergence

In this section we generalize two more ideas from single-variable analysis to general metric spaces. These are closed sets and convergent sequences.

Throughout this section let (X, d) and (Y, ρ) be metric spaces.

5.3.1 Closed Sets

Definition 5.3.1. A set $F \subset X$ is closed if it contains all of its limit points.

Theorem 5.3.2. A set $U \subset X$ is open if and only if its complement U^c is closed.

Proof. ($\implies$) Assume that U is open. For every $\mathbf{x} \in U$ there exists an $\varepsilon > 0$ such that $B(\mathbf{x}, \varepsilon) \subset U$, which implies that $B(\mathbf{x}, \varepsilon) \cap U^c$ is empty. Thus, no $\mathbf{x} \in U$ can be a limit point of U^c . In other words, U^c contains all its limit points. Therefore, U^c is closed.

($\impliedby$) Conversely, assume that U^c is closed and $\mathbf{x} \in U$. Since $\mathbf{x}$ cannot be a limit point of U^c , there exists $\varepsilon > 0$ such that $B(\mathbf{x}, \varepsilon) \subset U$. Thus, U is open. $\square$

Example 5.3.3. The disk $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$ is a closed subset of $\mathbb{R}^2$ because it contains all its limit points (see Example 5.2.14). Moreover, it is easy to see that its complement is open. The open ball $B(\mathbf{0}, 1) = \{(x, y) \mid x^2 + y^2 < 1\}$ is not closed because its complement is not open.

Nota Bene 5.3.4. Open and closed are not “opposite” properties. A set can be both open and closed at the same time. A set can also be neither open nor closed.

For example, we already know that $\emptyset$ and X are open, and by the previous theorem both $\emptyset$ and X are also closed sets, since $X = \emptyset^c$ and $\emptyset = X^c$. In the discrete metric every set is both open and closed. The interval $[0, 1] \subset \mathbb{R}$ in the usual metric is neither open nor closed.

Corollary 5.3.5. *The intersection of any collection of closed sets is closed, and the union of a finite collection of closed sets is closed.*

Proof. These follow from Theorems 5.1.14 and 5.3.2, via De Morgan’s laws (see Proposition A.1.11). $\square$

Remark 5.3.6. Note that the rules for intersection and union of closed set in Corollary 5.3.5 are the “opposite” of those for open sets given in Theorem 5.1.14.

Corollary 5.3.7. *Let (X, d) and (Y, ρ) be metric spaces. A function $f : X \rightarrow Y$ is continuous if and only if for each closed set $F \subset Y$ the preimage $f^{-1}(F)$ is closed in X .*

Proof. This follows from Theorem 5.2.3 and the fact that $f^{-1}(E^c) = f^{-1}(E)^c$; see Proposition A.2.4. $\square$

Example 5.3.8.* Here are more examples of closed sets.

- (i) The *closed ball centered at $\mathbf{x}_0$ with radius r* is the set $D(\mathbf{x}_0, r) = \{\mathbf{x} \in X \mid d(\mathbf{x}, \mathbf{x}_0) \leq r\}$. This set is closed by Corollary 5.3.7 because it is the preimage of the closed interval $[0, r]$ under the continuous map $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{x}_0)$; see Example 5.2.2(iv).
- (ii) Singleton sets are always closed, as are finite sets by Remark 5.2.20. Note that a singleton set $\{\mathbf{x}\}$ can also be written as the intersection of closed balls $\bigcap_{n=1}^{\infty} D(\mathbf{x}, \frac{1}{n}) = \{\mathbf{x}\}$.
- (iii) The unit circle $S^1 \subset \mathbb{R}^2$ is closed. Note that $f(x, y) = x^2 + y^2$ is continuous and the set $\{1\} \in \mathbb{R}$ is closed, so the set $f^{-1}(\{1\}) = \{(x, y) \mid x^2 + y^2 = 1\}$ is closed. In fact, for any continuous $f : X \rightarrow \mathbb{R}$, any set of the form $f^{-1}(\{c\}) \subset X$ is closed.^a

^aThe sets $f^{-1}(\{c\})$ are called *level sets* because if we consider the graph $\{(x, y, f(x, y)) \mid (x, y) \in \mathbb{R}^2\}$ of a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, then the set $f^{-1}(\{c\})$ is the set of all points of $\mathbb{R}^2$ that map to a point of height (level) c on the graph. Contour lines on a topographic map are level sets of the function that sends each point on the surface of the earth to its altitude.

5.3.2 The Closure of a Set

Definition 5.3.9. The closure of E , denoted $\overline{E}$, is the set E together with its limit points. We define the boundary of E , denoted ∂E , as the closure minus the interior, that is, $\partial E = \overline{E} \setminus E^\circ$.

Remark 5.3.10. A set $E \subset X$ is dense in X if and only if $\overline{E} = X$.

Theorem 5.3.11. The following properties hold for any $E \subset X$:

- (i) $\overline{E}$ is closed.
- (ii) If $F \subset X$ is closed and $E \subset F$, then $\overline{E} \subset F$; thus $\overline{E}$ is the smallest closed set containing E .
- (iii) E is closed if and only if $E = \overline{E}$.
- (iv) $\overline{E}$ is the intersection of all closed sets containing E .

Proof.

- (i) It suffices to show that $\overline{E}^c$ is open. If we denote the set of limit points of E by E' , then for any $\mathbf{p} \in \overline{E}^c = (E \cup E')^c = E^c \cap (E')^c$, there exists an $\varepsilon > 0$ such that $B(\mathbf{p}, \varepsilon) \setminus \{\mathbf{p}\} \subset E^c$. Combining this with the fact that $\mathbf{p} \in E^c$ gives $B(\mathbf{p}, \varepsilon) \subset E^c$. Moreover, if there exists $\mathbf{q} \in E' \cap B(\mathbf{p}, \varepsilon)$, then $B(\mathbf{p}, \varepsilon)$ is a neighborhood of $\mathbf{q}$ and therefore must contain a point of E , a contradiction. Therefore $B(\mathbf{p}, \varepsilon) \subset (E')^c$, which implies that $\overline{E}^c$ is open.
- (ii) It suffices to show that $E' \subset F'$. If $\mathbf{p} \in E'$, then for all $\varepsilon > 0$, we have that $B(\mathbf{p}, \varepsilon) \cap (E \setminus \{\mathbf{p}\}) \neq \emptyset$, which implies that $B(\mathbf{p}, \varepsilon) \cap (F \setminus \{\mathbf{p}\}) \neq \emptyset$ for all $\varepsilon > 0$. Thus, $\mathbf{p} \in F'$.
- (iii) If E is closed, then $\overline{E} \subset E$ by (ii), which implies that $\overline{E} = E$. Conversely, if $E = \overline{E}$, then E is closed by (i).
- (iv) Let F be the intersection of all closed sets containing E . By (ii), we have $\overline{E} \subset F$, since every closed set that contains E also contains $\overline{E}$. By (i), we have $F \subset \overline{E}$, since $\overline{E}$ is a closed set containing E . Thus, $\overline{E} = F$. $\square$

Example 5.3.12. Let $\mathbf{x} \in X$ and $r \geq 0$. Since $D(\mathbf{x}, r)$ is closed, we have that $\overline{B(\mathbf{x}, r)} \subset D(\mathbf{x}, r)$. If $X = \mathbb{F}^n$, then $\overline{B(\mathbf{x}, r)} = D(\mathbf{x}, r)$. However, in the discrete metric, we have that $\overline{B(\mathbf{x}, 1)} = \{\mathbf{x}\}$, whereas $D(\mathbf{x}, 1)$ is the entire space.

5.3.3 Sequences and Convergence

Sequences are a fundamental tool in analysis. They are most useful when they converge to a limit.

Definition 5.3.13. A sequence is a function $f : \mathbb{N} \rightarrow X$. We often write individual elements of the sequence as $\mathbf{x}_n = f(n)$ and the entire sequence as $(\mathbf{x}_i)_{i=0}^{\infty}$.

Definition 5.3.14. We say that $\mathbf{x} \in X$ is a limit of the sequence $(\mathbf{x}_i)_{i=0}^{\infty}$ if for all $\varepsilon > 0$ there exists $N > 0$ such that $d(\mathbf{x}, \mathbf{x}_n) < \varepsilon$ whenever $n \geq N$. We write $\mathbf{x}_k \rightarrow \mathbf{x}$ or $\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}$ and say that the sequence converges to $\mathbf{x}$.

Remark 5.3.15. By definition, a sequence $(\mathbf{x}_i)_{i=0}^{\infty}$ converges to $\mathbf{x}$ if and only if $\lim_{n \rightarrow \infty} d(\mathbf{x}_n, \mathbf{x}) = 0$. If the metric is defined in terms of a norm, then the sequence converges to $\mathbf{x}$ if and only if $\lim_{n \rightarrow \infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$.

Example 5.3.16.

- (i) Consider the sequence $((\frac{1}{n}, \frac{n}{n+1}))_{n=0}^{\infty}$ in the space $\mathbb{R}^2$. We prove that the sequence converges to the point $(0, 1)$. Given $\varepsilon > 0$, choose N so that $\frac{\sqrt{2}}{N} < \varepsilon$. Thus, whenever $n \geq N$, we have

$$\left\| (0, 1) - \left(\frac{1}{n}, \frac{n}{n+1} \right) \right\|_2 = \sqrt{\frac{1}{n^2} + \frac{1}{(n+1)^2}} < \frac{\sqrt{2}}{n} < \varepsilon.$$

- (ii)  Let $(f_n)_{n=0}^{\infty}$ be the sequence of functions in $C([0, 1]; \mathbb{R})$ given by $f_n(t) = t^{1/n}$, and let f be the constant function $f(t) = 1$. We show that $\|f - f_n\|_{L^2} \rightarrow 0$, so in the L^2 -norm we have $f_n \rightarrow f$; but $\|f - f_n\|_{L^\infty} = 1$ for all $n \in \mathbb{N}$, and so f_n does not converge to f in the L^∞ -norm.

As $n \rightarrow \infty$, we have

$$\|f - f_n\|_{L^2}^2 = \int_0^1 |1 - t^{1/n}|^2 dt = \frac{2}{(n+1)(n+2)} \rightarrow 0.$$

In contrast, we also have

$$\|f - f_n\|_{L^\infty} = \sup_{t \in [0, 1]} |f(t) - f_n(t)| = \sup_{t \in [0, 1]} |1 - t^{1/n}| \rightarrow 1.$$

Proposition 5.3.17. If a sequence has a limit, it is unique.

Proof. Let $(\mathbf{x}_n)_{n=0}^{\infty}$ be a sequence in X . Suppose that $\mathbf{x}_n \rightarrow \mathbf{x}$ and also that $\mathbf{x}_n \rightarrow \mathbf{y} \neq \mathbf{x}$. For $\varepsilon = d(\mathbf{x}, \mathbf{y})$, there exists $N > 0$ so that $d(\mathbf{x}_n, \mathbf{x}) < \frac{\varepsilon}{2}$ whenever $n \geq N$. Similarly, there exists $m > N$ with $d(\mathbf{x}_m, \mathbf{y}) < \frac{\varepsilon}{2}$. However, this implies $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{x}_m) + d(\mathbf{x}_m, \mathbf{y}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon = d(\mathbf{x}, \mathbf{y})$, which is a contradiction. $\square$

Nota Bene 5.3.18. Limits and limit points are fundamentally different things. Sequences have limits, and sets have limit points. The limit of a sequence, if it exists at all, is unique but need not be a limit point of the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ of terms in the sequence. Conversely, a limit point of the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ of terms in the sequence need not be the limit of the sequence. The examples below illustrate these differences.

(i) The sequence

$$\mathbf{x}_n = \begin{cases} 1/n & \text{if } n \text{ even,} \\ 1 - 1/n & \text{if } n \text{ odd} \end{cases}$$

does not converge to any limit. Both of the points 1 and 0 are limit points of the set $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots\}$.

(ii) The sequence $2, 2, 2, \dots$ (that is, the sequence $\mathbf{x}_n = 2$ for every n) converges to the limit 2, but the set $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots\} = \{2\}$ has no limit points.

Theorem 5.3.19. A function $f : X \rightarrow Y$ is continuous at a point $\mathbf{x}_* \in X$ if and only if, for each sequence $(\mathbf{x}_k)_{k=0}^\infty \subset X$ that converges to $\mathbf{x}_*$, the sequence $(f(\mathbf{x}_k))_{k=0}^\infty \subset Y$ converges to $f(\mathbf{x}_*) \in Y$.

Proof. ($\Rightarrow$) Assume that f is continuous at $\mathbf{x}_* \in X$. Thus, given $\varepsilon > 0$ there exists a $\delta > 0$ such that $\rho(f(\mathbf{x}), f(\mathbf{x}_*)) < \varepsilon$ whenever $d(\mathbf{x}, \mathbf{x}_*) < \delta$. Since $\mathbf{x}_k \rightarrow \mathbf{x}_*$, there exists $N > 0$ such that $d(\mathbf{x}_n, \mathbf{x}_*) < \delta$ whenever $n \geq N$. Thus, $\rho(f(\mathbf{x}_n), f(\mathbf{x}_*)) < \varepsilon$ whenever $n \geq N$, and therefore $(f(\mathbf{x}_k))_{k=0}^\infty$ converges to $f(\mathbf{x}_*)$.

($\Leftarrow$) If f is not continuous at $\mathbf{x}_* \in X$, then there exists $\varepsilon > 0$ such that for each $\delta > 0$ there exists $\mathbf{x} \in B(\mathbf{x}_*, \delta)$ with $\rho(f(\mathbf{x}), f(\mathbf{x}_*)) \geq \varepsilon$. For each $k \in \mathbb{Z}^+$, choose $\mathbf{x}_k \in B(\mathbf{x}_*, \frac{1}{k})$ with $\rho(f(\mathbf{x}_k), f(\mathbf{x}_*)) \geq \varepsilon$. This implies that $\mathbf{x}_k \rightarrow \mathbf{x}_*$, but $(f(\mathbf{x}_k))_{k=0}^\infty$ does not converge to $f(\mathbf{x}_*)$ because $\rho(f(\mathbf{x}_k), f(\mathbf{x}_*)) \geq \varepsilon$. $\square$

Example 5.3.20. We can often use continuous functions to prove that a sequence converges. For example, the sequence $x_n = \sin(\frac{1}{n})$ converges to zero, since $\frac{1}{n} \rightarrow 0$ as $n \rightarrow \infty$ and the sine function is continuous and equal to zero at zero.

Unexample 5.3.21.

(i) The function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

$$f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0), \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

is not continuous at the origin because if $\mathbf{x}_n = (1/n, 1/n)$ for every $n \in \mathbb{Z}^+$, we have $f(\mathbf{x}_n) = 1/2$ for every n , but $f(\lim_{n \rightarrow \infty} \mathbf{x}_n) = 0$.

- (ii)  The derivative map $D(f) = f'$ on the set $\mathbb{F}[x]$ of polynomials is not continuous in the L^∞ -norm on $C([0, 1]; \mathbb{F})$. For each $n \in \mathbb{Z}^+$ let $f_n(x) = \frac{x^n}{n}$. Note that $\|f_n\|_{L^\infty} = \frac{1}{n}$; therefore, $f_n \rightarrow 0$. And yet $\|D(f_n)\|_{L^\infty} = 1$, so $D(f_n)$ does not converge to $D(0) = 0$. Since D does not preserve limits, it is not continuous at the origin. See Exercise 5.21 for a generalization.

Corollary 5.3.22. *For any $\mathbf{x}, \mathbf{y} \in X$ and any sequence $(\mathbf{x}_n)_{n=0}^\infty$ in X converging to $\mathbf{x}$, we have*

$$\lim_{n \rightarrow \infty} d(\mathbf{x}_n, \mathbf{y}) = d(\lim_{n \rightarrow \infty} \mathbf{x}_n, \mathbf{y}) = d(\mathbf{x}, \mathbf{y}),$$

or, equivalently, if the metric is $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ for some norm $\|\cdot\|$, then

$$\lim_{n \rightarrow \infty} \|\mathbf{x}_n - \mathbf{y}\| = \|\lim_{n \rightarrow \infty} \mathbf{x}_n - \mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|. \quad (5.6)$$

Proof. The function $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{y})$ is continuous (see Example 5.2.2(iv)), so the result follows from Theorem 5.3.19. $\square$

5.4 Completeness and Uniform Continuity

Convergent sequences are very important, but in order to use the definition of convergence, you need to know the limit of the sequence. A useful way to get around this problem is to observe that all convergent sequences are also *Cauchy sequences*. In many situations this allows us to identify convergent sequences without knowing their limits. In some sense, Cauchy sequences are those that *should* converge, even if they have no limit. Metric spaces in which all Cauchy sequences converge are especially useful—these are called *complete* spaces.

Unfortunately, not all continuous functions preserve Cauchy sequences—some continuous functions map some Cauchy sequences into sequences that are not Cauchy. So we need a stronger form of continuity called *uniform continuity*.

In this section we define and discuss Cauchy sequences, completeness, and uniform continuity. We also show that uniformly continuous functions preserve Cauchy sequences, whereas continuous functions may not. We conclude the section by proving that $\mathbb{F}^n$ is complete.

Throughout this section assume that (X, d) and (Y, ρ) are metric spaces.

5.4.1 Cauchy Sequences and Completeness

Definition 5.4.1. *A sequence $(\mathbf{x}_i)_{i=1}^\infty$ in X is a Cauchy sequence if for all $\varepsilon > 0$ there exists an $N > 0$ such that $d(\mathbf{x}_m, \mathbf{x}_n) < \varepsilon$ whenever $m, n \geq N$.*

Example 5.4.2.

- (i) The sequence $(x_n)_{n=1}^{\infty}$ given by $x_n = 1/n$ in $\mathbb{R}$ is a Cauchy sequence: given $\varepsilon > 0$, choose $N > 1/\varepsilon$. If $n \geq m \geq N$, then

$$|1/m - 1/n| = |(n - m)/mn| < |n/mn| = |1/m| \leq 1/N < \varepsilon.$$

- (ii) Fix $b \in (0, 1) \subset \mathbb{R}$. The sequence $(f_n(x))_{n=0}^{\infty}$ given by $f_n(x) = x^n$ is Cauchy in the space $C([0, b]; \mathbb{R})$ of continuous functions on $[0, b]$ with the metric induced by the sup norm. To see this, for any $\varepsilon > 0$ let $N > \log_b(\varepsilon)$. If $n \geq m \geq N$, then for every $x \in [0, b]$ we have $|f_m(x) - f_n(x)| \leq |x^m| \leq b^m < \varepsilon$.

Unexample 5.4.3.

- (i) Even when the difference $x_n - x_{n-1}$ goes to zero, the sequence $(x_n)_{n=1}^{\infty}$ need not be Cauchy. For example, the sequence given by $x_n = \log(n)$ in $\mathbb{R}$ satisfies $x_n - x_{n-1} = \log(n/(n-1)) = \log(1 + 1/(n-1)) \rightarrow 0$, but the sequence is not Cauchy because for any m we may take $n = km$ and then $|x_n - x_m| = \log(n/m) = \log(k)$. This difference can be made arbitrarily large, and thus does not satisfy the Cauchy criterion.
- (ii) The sequence given by $f_n(x) = nx^n$ is *not* a Cauchy sequence in the space $C([0, 1]; \mathbb{R})$ of continuous functions on $[0, 1]$ with the metric induced by the sup norm. To see this, we note that

$$\|f_m - f_n\|_{L^\infty} = \sup_{x \in [0, 1]} |f_m - f_n| \geq |f_m(1) - f_n(1)| = |m - n|,$$

which cannot be made small by requiring n and m to be large.

Proposition 5.4.4. *Any sequence that converges is a Cauchy sequence.*

Proof. Assume $(\mathbf{x}_i)_{i=0}^{\infty}$ in X converges to some $\mathbf{x} \in X$. Given $\varepsilon > 0$, there exists an $N > 0$ such that $d(\mathbf{x}_n, \mathbf{x}) < \varepsilon/2$ whenever $n \geq N$. Hence, we have

$$d(\mathbf{x}_m, \mathbf{x}_n) \leq d(\mathbf{x}_m, \mathbf{x}) + d(\mathbf{x}_n, \mathbf{x}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

whenever $m, n \geq N$. $\square$

Nota Bene 5.4.5. Not all Cauchy sequences are convergent. For example, in the space $X = \mathbb{R} \setminus \{0\}$ with the usual metric $d(x, y) = |x - y|$, the sequence $(1/n)_{n=1}^{\infty}$ is Cauchy, and it does converge in $\mathbb{R}$, but it does not converge in X because its limit 0 is not in X .

Similarly, in the space $\mathbb{Q}$, the sequence $((1 + 1/n)^n)_{n=1}^\infty$ is Cauchy and it converges in $\mathbb{R}$ to e , but it does not converge in $\mathbb{Q}$ because $e \notin \mathbb{Q}$.

We show later that all Cauchy sequences in $\mathbb{R}$ converge, but there are many other important spaces where not all Cauchy sequences converge.

Definition 5.4.6. We say that a set $E \subset X$ is bounded if for every $\mathbf{x} \in X$ there is a positive real number M such that $d(\mathbf{p}, \mathbf{x}) < M$ for every $\mathbf{p} \in E$.

Proposition 5.4.7. Cauchy sequences are bounded.

Proof. Let $(\mathbf{x}_k)_{k=0}^\infty$ be a Cauchy sequence, and choose $\varepsilon = 1$. Thus, there exists $N > 0$ such that $d(\mathbf{x}_n, \mathbf{x}_m) < 1$ whenever $m, n \geq N$. Hence, for any fixed $\mathbf{x} \in X$ and any $m \geq N$, we have that

$$d(\mathbf{x}_n, \mathbf{x}) \leq d(\mathbf{x}_n, \mathbf{x}_m) + d(\mathbf{x}_m, \mathbf{x}) \leq 1 + d(\mathbf{x}_m, \mathbf{x}),$$

whenever $n \geq N$. Setting

$$M = \max\{d(\mathbf{x}_1, \mathbf{x}), \dots, d(\mathbf{x}_{N-1}, \mathbf{x}), 1 + d(\mathbf{x}_m, \mathbf{x})\}$$

gives $d(\mathbf{x}_k, \mathbf{x}) \leq M$ for all $k \in \mathbb{N}$. $\square$

We now need the idea of a subsequence, which is, as the name suggests, a sequence consisting of some, but not necessarily all, the elements of the original sequence. Here is a careful definition.

Definition 5.4.8. A subsequence of a sequence $(\mathbf{x}_n)_{n=0}^\infty$ is a sequence of the form $(\mathbf{x}_{n_i})_{i=0}^\infty$, where $n_0 < n_1 < n_2 < \dots$ are nonnegative integers.

Proposition 5.4.9. Any Cauchy sequence that has a convergent subsequence is convergent.

Proof. Let $(\mathbf{x}_n)_{n=0}^\infty$ be a Cauchy sequence. If $(\mathbf{x}_{n_j})_{j=1}^\infty$ is a subsequence that converges to $\mathbf{x}$, then we claim that $(\mathbf{x}_n)_{n=0}^\infty$ converges to $\mathbf{x}$. To see this, for any $\varepsilon > 0$ choose N such that $\|\mathbf{x}_n - \mathbf{x}_m\| \leq \varepsilon/2$ whenever $m, n > N$, and choose J such that $\|\mathbf{x}_{n_j} - \mathbf{x}\| \leq \varepsilon/2$ whenever $j > J$. For any $m > N$, choose a $j > J$ such that $n_j > N$. We have

$$\|\mathbf{x}_m - \mathbf{x}\| \leq \|\mathbf{x}_m - \mathbf{x}_{n_j}\| + \|\mathbf{x}_{n_j} - \mathbf{x}\| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Hence, the sequence converges to $\mathbf{x}$. $\square$

Example 5.4.10. Convergence is preserved under continuity, but Cauchy sequences may not be. As an example, let $h : [0, 1) \rightarrow \mathbb{R}$ be defined by $h(x) = \frac{x}{1-x}$. Note that $x_n = 1 - \frac{1}{n}$ is Cauchy, but $h(x_n)$ is not, since $h(x_n) = n - 1$.

We show in the next subsection that a stronger version of continuity called *uniform continuity* does preserve Cauchy sequences.

Nota Bene 5.4.5 gives some examples of Cauchy sequences that have no limit. In some sense, these indicate a hole or gap in the space, leaving it incomplete. This motivates the following definition.

Definition 5.4.11. A metric space (X, d) is complete if every Cauchy sequence converges.

Unexample 5.4.12.

- (i) The set $\mathbb{Q}$ of rational numbers with the usual metric $d(x, y) = |x - y|$ is not complete. For example, let $(x_n)_{n=0}^\infty$ be the sequence

$$3, 3.1, 3.14, 3.141, 3.1415, \dots,$$

with x_n consisting of the decimal approximation of π to the first $n + 1$ places. This sequence converges to π in $\mathbb{R}$, so it is Cauchy, but it has no limit in $\mathbb{Q}$.

- (ii) The space $\mathbb{R} \setminus \{0\}$ is not complete (see Nota Bene 5.4.5).

- (iii) The vector space $C([0, 2]; \mathbb{R})$ with the L^1 -norm $\|f\|_{L^1} = \int_0^2 |f(t)| dt$ is not complete. To see this, consider the sequence $(g_n)_{n=0}^\infty$ defined by

$$g_n(t) = \begin{cases} t^n, & t \leq 1, \\ 1, & t \geq 1. \end{cases}$$

Given $\varepsilon > 0$ let $N \geq 1/\varepsilon$. Since every g_n is equal to every g_m on the interval $[1, 2]$, if $m, n > N$, then

$$\begin{aligned} \|g_n - g_m\|_{L^1} &= \int_0^1 |t^n - t^m| dt \\ &= |1/(n+1) - 1/(m+1)| \\ &= \frac{|m-n|}{(m+1)(n+1)} < \varepsilon. \end{aligned}$$

Therefore, this sequence is Cauchy in the L^1 -norm, but $(g_n)_{n=0}^\infty$ does not converge to any continuous function in the L^1 -norm. In fact, in the L^1 -norm we have $g_n \rightarrow g$ where

$$g = \begin{cases} 0, & t < 1, \\ 1, & t \geq 1. \end{cases}$$

We do not prove this here, but it follows easily from the monotone convergence theorem (Theorem 8.4.5).

Remark 5.4.13. In Section 9.1.2 we show that every metric space can be uniquely extended to a complete metric space by creating equivalence classes of Cauchy

sequences; that is, two Cauchy sequences are equivalent if the distance between them converges to zero. This is also the idea behind one construction of the real numbers $\mathbb{R}$ from $\mathbb{Q}$. In other words, $\mathbb{R}$ can be constructed as the set of all Cauchy sequences in $\mathbb{Q}$ modulo this equivalence relation (see also Vista 9.1.4).

Theorem 5.4.14. *The fields $\mathbb{R}$ and $\mathbb{C}$ are complete with respect to the usual metric $d(x, y) = |x - y|$.*

We prove this theorem and the next one in Section 5.4.3.

Theorem 5.4.15. *If $((X_i, d_i))_{i=1}^n$ is a finite collection of complete metric spaces, then the Cartesian product $X = X_1 \times X_2 \times \cdots \times X_n$ is complete when endowed with the p -metric (5.4) for $1 \leq p \leq \infty$.*

The following theorem is fundamental and is an immediate corollary of the previous two theorems.

Theorem 5.4.16. *For every $n \in \mathbb{N}$ and $p \in [1, \infty]$, the linear space $\mathbb{F}^n$ with the norm $\|\cdot\|_p$ is complete.*

Remark 5.4.17. While the previous theorem shows that $\mathbb{F}^n$ is complete in the p -metric, Corollary 5.4.25 shows that $\mathbb{F}^n$ is complete in *any* metric that is induced by a norm (but not necessarily in metrics that are not induced by a norm, like the discrete metric).

5.4.2 Uniform Continuity

Continuous functions are fundamental in analysis, but sometimes continuity is not enough. For example, continuous functions don't preserve Cauchy sequences. What we need in these cases is a stronger condition called *uniform continuity*. You may have encountered this in single-variable analysis, and it has a natural generalization to arbitrary metric spaces.

Uniform continuity requires that there be no reference point for the continuity—the relationship between δ and ε must be independent of the point in question.

Definition 5.4.18. *A function $f : X \rightarrow Y$ is uniformly continuous on $E \subset X$ if for all $\varepsilon > 0$ there exists $\delta > 0$ such that $\rho(f(\mathbf{x}), f(\mathbf{y})) < \varepsilon$ whenever $\mathbf{x}, \mathbf{y} \in E$ and $d(\mathbf{x}, \mathbf{y}) < \delta$.*

Example 5.4.19. The function $f(x) = e^x$ is uniformly continuous on any closed interval $[a, b]$. To see this, note that e^x is continuous, so for any $\varepsilon > 0$ there is a $\delta > 0$ such that $|1 - e^z| < \varepsilon/e^b$ if $|z| < \delta$. Thus, for any $x, y \in [a, b]$ with $x > y$ and $|x - y| < \delta$, we have $|e^x - e^y| = |1 - e^{y-x}||e^x| < (\varepsilon/e^b)|e^x| < \varepsilon$, as required.

Unexample 5.4.20. The function $f(x) = e^x$ is not uniformly continuous on $\mathbb{R}$ because for any fixed difference $|x - y|$, no matter how small, we can make $|e^x - e^y| = |e^x||1 - e^{y-x}|$ as large as desired by making x (and hence e^x) as large as needed.

The function $g(x) = 1/x$ is not uniformly continuous on $(0, 1)$ since for any fixed value of $|x - y|$, the difference $|1/x - 1/y| = |x - y|/|xy|$ can be made as large as desired by making $|xy|$ sufficiently small.

Roughly speaking, the functions f and g in this unexample curve upward more and more as they approach one edge of the domain, and this is what causes uniform continuity to fail for them.

Example 5.4.21. A Lipschitz continuous function $f : X \rightarrow Y$ with constant K (see Example 5.2.2(iii)) is uniformly continuous on X . If $\rho(f(\mathbf{x}), f(\mathbf{y})) \leq Kd(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in X$, then, given $\varepsilon > 0$, choose $\delta = \frac{\varepsilon}{K}$. Thus, $\rho(f(\mathbf{x}), f(\mathbf{y})) < \varepsilon$ whenever $d(\mathbf{x}, \mathbf{y}) < \delta$.

An important class of uniformly continuous functions are bounded linear transformations.

Proposition 5.4.22. *If $f : X \rightarrow Y$ is a bounded linear transformation of normed linear spaces, then f is uniformly continuous. Conversely, any continuous linear transformation is bounded.*

Proof. The proof that bounded linear transformations are uniformly continuous is Exercise 5.24. Conversely, Exercise 5.21 shows that any continuous linear transformation is bounded. $\square$

Remark 5.4.23. Although not every continuous function is uniformly continuous, we prove in the next section (Theorem 5.5.9) that if the domain is compact (or if it is complete and totally bounded), then a continuous function must, in fact, be uniformly continuous.

Recall that Cauchy sequences are not preserved under continuity (see Example 5.4.10). The following theorem says that they are, however, preserved under uniform continuity.

Theorem 5.4.24. *Assume $f : X \rightarrow Y$ is uniformly continuous. If $(\mathbf{x}_k)_{k=0}^\infty$ is a Cauchy sequence, then so is $(f(\mathbf{x}_k))_{k=0}^\infty$.*

Proof. Let $\varepsilon > 0$ be given. Since f is uniformly continuous on X , there exists a $\delta > 0$ such that $\rho(f(\mathbf{x}), f(\mathbf{y})) < \varepsilon$ whenever $d(\mathbf{x}, \mathbf{y}) < \delta$. Since $(\mathbf{x}_k)_{k=0}^\infty$ is Cauchy, there exists $N > 0$ such that $d(\mathbf{x}_m, \mathbf{x}_n) < \delta$ whenever $n, m \geq N$. Thus, $\rho(f(\mathbf{x}_m), f(\mathbf{x}_n)) < \varepsilon$ whenever $m, n \geq N$, and so $(f(\mathbf{x}_k))_{k=1}^\infty$ is Cauchy. $\square$

Corollary 5.4.25. *Every finite-dimensional normed linear space over $\mathbb{F}$ is complete.*

Proof. Given any finite-dimensional normed linear space $(Z, \|\cdot\|)$, Corollary 2.3.12 guarantees there is an isomorphism of vector spaces $f : Z \rightarrow \mathbb{F}^n$. We can make $\mathbb{F}^n$ into a normed linear space with the Euclidean norm $\|\cdot\|_2$. By Remark 3.5.13, every linear transformation of finite-dimensional normed linear spaces is bounded, and thus f and f^{-1} are both bounded linear transformations. Moreover, Proposition 5.4.22 guarantees that bounded linear transformations are uniformly continuous.

Given any Cauchy sequence $(\mathbf{z}_k)_{k=0}^\infty$ in Z , for each k let $\mathbf{y}_k = f(\mathbf{z}_k) \in \mathbb{F}^n$. By Theorem 5.4.24, the sequence $(\mathbf{y}_k)_{k=0}^\infty$ must also be Cauchy, and thus has a limit $\mathbf{y} \in \mathbb{F}^n$, since $(\mathbb{F}^n, \|\cdot\|_2)$ is complete. For each k we have $f^{-1}(\mathbf{y}_k) = \mathbf{z}_k$, and so $\lim_{k \rightarrow \infty} \mathbf{z}_k = \lim_{k \rightarrow \infty} f^{-1}(\mathbf{y}_k) = f^{-1}(\mathbf{y})$ exists, since f^{-1} is continuous. $\square$

Finally, we conclude with a lemma that is important when we talk about integration in Section 5.10 and again in Chapter 8.

Lemma 5.4.26. *If Y is a dense subspace of a normed linear space Z such that every Cauchy sequence in Y converges in Z , then Z is complete.*

Proof. The proof is Exercise 5.25. $\square$

5.4.3 * Proof that $\mathbb{F}^n$ is complete

In this section we prove the very important Theorem 5.4.16, which tells us $\mathbb{F}^n$ is complete with respect to the metric induced by the p -norm for any $p \in [1, \infty]$. To do this we begin with a review from single-variable calculus of the proof that $\mathbb{R}$ is complete. Using the fact that $\mathbb{R}$ is complete, it is straightforward to show that $\mathbb{C}$ is also complete. We then prove Theorem 5.4.15 that Cartesian products of complete spaces with the p -metric are complete. Combining this with the previous results gives Theorem 5.4.16.

The single-variable case

The completeness of $\mathbb{R}$ relies upon the following fundamental property of the real numbers called *Dedekind's property* or the *least upper bound property*. For a proof of this property, we refer the reader to [HS75].

Theorem 5.4.27. *Every nonempty subset of the real numbers that is bounded above has a supremum.*

To begin the proof of completeness, we need a lemma.

Lemma 5.4.28. *Every sequence $(y_n)_{n=0}^\infty$ in $\mathbb{R}$ that is bounded above and is monotone increasing (that is, $y_n \leq y_{n+1}$ for every $n \in \mathbb{N}$) has a limit.*

Proof. Since the set $\{y_n \mid n \in \mathbb{N}\}$ is bounded above, it has a supremum (least upper bound) by Dedekind's property. Let $y = \sup\{y_n \mid n \in \mathbb{N}\}$. If $\varepsilon > 0$, then

there exists some $m > 0$ such that $|y_m - y| < \varepsilon$. If not, then $y - \varepsilon/2$ would also be an upper bound for $\{y_n \mid n \in \mathbb{N}\}$ and y would not be the supremum.

Since the sequence is monotone increasing, we must have $y_m \leq y_n \leq y$ for every $n > m$. Therefore, $|y - y_n| \leq |y - y_m| < \varepsilon$ for all $n > m$, so y is the desired limit. $\square$

Corollary 5.4.29. *Every sequence in $\mathbb{R}$ that is bounded below and is monotone decreasing has a limit.*

Proof. Let $(z_n)_{n=0}^\infty$ be a monotone decreasing sequence that is bounded below by B . Let $y_n = -z_n$ for every $n \in \mathbb{N}$. The sequence $(y_n)_{n=0}^\infty$ is monotone increasing and bounded above by $-B$, and so it converges to a limit y .

The sequence $(z_n)_{n=0}^\infty$ converges to $z = -y$ because for every $\varepsilon > 0$, if we choose N such that $|y - y_n| < \varepsilon$ for all $n > N$, then $|z - z_n| = |-y - (-y_n)| = |y - y_n| < \varepsilon$. $\square$

Theorem 5.4.30. *The space $\mathbb{R}$ with the usual metric $d(x, y) = |x - y|$ is complete.*

Proof. Let $(x_n)_{n=0}^\infty$ be a Cauchy sequence in $\mathbb{R}$. Since every Cauchy sequence is bounded, for each $n \in \mathbb{N}$ the set $E_n = \{x_i \mid i \geq n\}$ is bounded, and hence has a supremum. For each n , let $y_n = \sup E_n$. The sequence $(y_n)_{n=0}^\infty$ is monotonically decreasing and is bounded, so by Corollary 5.4.29 it converges to some value y .

For any $\varepsilon > 0$, choose N such that $|x_m - x_n| \leq \varepsilon/3$ whenever $n, m > N$. Choose $k > N$ such that $|y - y_k| < \varepsilon/3$. Since y_k is the supremum of E_k , there must be some $m \geq k$ such that $|y_k - x_m| < \varepsilon/3$. For every $n \geq N$ we have

$$|y - x_n| \leq |y - y_k| + |y_k - x_m| + |x_m - x_n| < \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \quad \square$$

Proposition 5.4.31. *The space $\mathbb{C}$ with the usual metric $d(z, w) = |z - w|$ is complete.*

Proof. The proof is Exercise 5.28. $\square$

Higher dimensions

Now we prove Theorem 5.4.15, which says that if $((X_i, d_i))_{i=1}^n$ is a finite collection of complete metric spaces, then the Cartesian product $X = X_1 \times X_2 \times \cdots \times X_n$ is complete when endowed with the p -metric (5.4) for $1 \leq p \leq \infty$.

Proof. Let $(\mathbf{x}_k)_{k=0}^\infty \subset X$ be a Cauchy sequence. Denote the components of $\mathbf{x}_k$ as $\mathbf{x}_k = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})$ with $x_i^{(k)} \in X_i$ for each i . For every $\varepsilon > 0$, choose an N such that for every $\ell, m > N$ we have $\varepsilon > d^p(\mathbf{x}_\ell, \mathbf{x}_m)$. If $p < \infty$, then we have

$$\varepsilon > d^p(\mathbf{x}_\ell, \mathbf{x}_m) = \left(\sum_{i=1}^n d_i(x_i^{(\ell)}, x_i^{(m)})^p \right)^{1/p} \geq \left(d_j(x_j^{(\ell)}, x_j^{(m)})^p \right)^{1/p} = d_j(x_j^{(\ell)}, x_j^{(m)})$$

for every j . If $p = \infty$, then we have

$$\varepsilon > d^\infty(\mathbf{x}_\ell, \mathbf{x}_m) = \sup_i d_i(x_i^{(\ell)}, x_i^{(m)}) \geq d_j(x_j^{(\ell)}, x_j^{(m)})$$

for every j . In either case, each sequence $(x_j^{(k)})_{k=0}^\infty \subset X_j$ is Cauchy and converges to some $x_j \in X_j$.

Define $\mathbf{x} = (x_1, \dots, x_n)$. We show that $(\mathbf{x}_k)_{k=0}^\infty$ converges to $\mathbf{x}$. Given $\varepsilon > 0$, choose N so that

$$d_i(x_i^{(\ell)}, x_i^{(m)}) \leq d^p(\mathbf{x}_\ell, \mathbf{x}_m) < \frac{\varepsilon}{n+1},$$

whenever $\ell, m \geq N$. Letting $\ell \rightarrow \infty$ gives $d_i(x_i, x_i^{(m)}) \leq \frac{\varepsilon}{n+1}$. When $p = \infty$, this gives $d^\infty(\mathbf{x}, \mathbf{x}_m) \leq \frac{\varepsilon}{n+1} < \varepsilon$, and when $p \in [1, \infty)$, this gives

$$d^p(\mathbf{x}, \mathbf{x}_m) \leq \left(\sum_{i=1}^n \left(\frac{\varepsilon}{n+1} \right)^p \right)^{1/p} \leq n \left(\frac{\varepsilon}{n+1} \right) < \varepsilon.$$

The next-to-last inequality follows from the fact that $a^p + b^p \leq (a+b)^p$ for any nonnegative numbers a, b and any $p \in [1, \infty)$.

For all p we now have $d^p(\mathbf{x}, \mathbf{x}_m) < \varepsilon$, whenever $m \leq N$. It follows that the Cauchy sequence $(\mathbf{x}_k)_{k=0}^\infty \subset X$ converges and that X is complete. $\square$

Nota Bene 5.4.32. The previous proof uses a technique often seen in analysis, especially when working with a convergent sequence $\mathbf{x}_k \rightarrow \mathbf{x}$. The technique is to bound $d(\mathbf{x}_\ell, \mathbf{x}_m)$ and then let $\ell \rightarrow \infty$ to get a bound on $d(\mathbf{x}, \mathbf{x}_m)$. This works because the function $d(\cdot, \mathbf{x}_m)$ is continuous (see Example 5.2.2(iv)).

5.5 Compactness

Recall from Corollary 5.3.7 that for a continuous function $f : X \rightarrow Y$ and for a closed subset $Z \subset Y$, the preimage $f^{-1}(Z)$ is closed. But if $W \subset X$ is closed, then the image $f(W)$ is not necessarily closed. In this section we discuss a property of a set called *compactness*, which seems just slightly stronger than being closed, but which has many powerful consequences. One of these is that continuous functions map compact sets to compact sets, and this guarantees that every real-valued continuous function on a compact set attains both its minimum and its maximum. These lead to many other important consequences that you will use in this book and far beyond.

The definition of a compact set may seem somewhat strange—it is given in unfamiliar terms involving various collections of open sets—but we prove the *Heine–Borel theorem* (see Theorem 5.5.4 and also 5.5.12), which gives a simple description of compact sets in $\mathbb{R}^n$ as those that are both closed and bounded.

Throughout this section we assume that (X, d) and (Y, ρ) are metric spaces.

5.5.1 Introduction to Compactness

Definition 5.5.1. A collection $(G_\alpha)_{\alpha \in J}$ of open sets is an *open cover* of the set E if $E \subset \bigcup_{\alpha \in J} G_\alpha$. A set E is *compact* if every open cover has a *finite subcover*; that is, for every open cover $(G_\alpha)_{\alpha \in J}$ there exists a *finite subcollection* $(G_\alpha)_{\alpha \in J'}$, where $J' \subset J$ is a finite subset, such that $E \subset \bigcup_{\alpha \in J'} G_\alpha$.

Proposition 5.5.2. *A closed subset of a compact set is compact.*

Proof. Let F be a closed subset of a compact set K , and let $\mathcal{C} = (G_\alpha)_{\alpha \in J}$ be an open covering of F . Thus, $\mathcal{C} \cup \{F^c\}$ is an open covering of K , which has a finite subcovering $\{F^c, G_{\alpha_1}, \dots, G_{\alpha_n}\}$. Hence, $(G_{\alpha_k})_{k=1}^n$ is a finite subcover of F . $\square$

Theorem 5.5.3. *A compact subset of a metric space is closed and bounded.*

Proof. Fix a point $\mathbf{x} \in X$. For any compact subset $K \subset X$, the collection $(B(\mathbf{x}, k))_{k=1}^\infty$ is an open cover of K . It must have a finite subcover $(B(\mathbf{x}, k))_{k=1}^M$, since K is compact. Therefore $K \subset B(\mathbf{x}, M)$, which implies that K is bounded.

To see that K is closed, we may assume that $K \neq \emptyset$ and $K \neq X$ because these two sets are closed. We show that K^c is open. Assume that $\mathbf{x} \in K^c$. For every $\mathbf{y} \in K$ let $\delta_{\mathbf{y}} = d(\mathbf{x}, \mathbf{y})/2$. The collection of balls $(B(\mathbf{y}, \delta_{\mathbf{y}}))_{\mathbf{y} \in K}$ forms an open cover of K , so there exists a finite subcover $B(\mathbf{y}_1, \delta_{\mathbf{y}_1}), \dots, B(\mathbf{y}_n, \delta_{\mathbf{y}_n})$. Let $\delta = \min(\delta_{\mathbf{y}_1}, \dots, \delta_{\mathbf{y}_n})$. This is strictly positive, and we claim that $B(\mathbf{x}, \delta) \cap K = \emptyset$. To see this, observe that for any $\mathbf{z} \in K$, we have $\mathbf{z} \in B(\mathbf{y}_i, \delta_{\mathbf{y}_i})$ for some i . Thus, by a variation on the triangle inequality (see Exercise 5.5), we have $d(\mathbf{z}, \mathbf{x}) \geq d(\mathbf{x}, \mathbf{y}_i) - d(\mathbf{z}, \mathbf{y}_i) \geq d(\mathbf{x}, \mathbf{y}_i)/2 \geq \delta$. Hence, K^c is open, and K is closed. $\square$

The next theorem gives a partial converse to Theorem 5.5.3.

Theorem 5.5.4 (Heine–Borel). *If a subset of $\mathbb{R}^n$ (with the usual, Euclidean metric) is closed and bounded, then it is compact.*

Proof. First we show that every n -cell is compact, that is, every set of the form $[\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$ (meaning $a_k \leq x_k \leq b_k$ for all $k = 1, \dots, n$).

Suppose $(G_\alpha)_{\alpha \in J}$ is an open cover of $I_1 = [\mathbf{a}, \mathbf{b}]$ that contains no finite subcover. Let $\mathbf{c} = \frac{\mathbf{a}+\mathbf{b}}{2}$ be the midpoint of $\mathbf{a}$ and $\mathbf{b}$, meaning each $c_k = \frac{a_k+b_k}{2}$. The intervals $[a_k, c_k]$ and $[c_k, b_k]$ determine 2^n n -cells, at least one of which, denoted I_2 , cannot be covered by a finite subcollection of $(G_\alpha)_{\alpha \in J}$. Subdivide I_2 and repeat. We have a sequence $(I_k)_{k=1}^\infty$ of n -cells such that $I_{n+1} \subset I_n$, where each I_n is not covered by any finite subcollection of $(G_\alpha)_{\alpha \in J}$ and $\mathbf{x}, \mathbf{y} \in I_n$ implies $\|\mathbf{x} - \mathbf{y}\|_2 \leq 2^{-n} \|\mathbf{b} - \mathbf{a}\|_2$.

By choosing $\mathbf{x}_k \in I_k$, we have a Cauchy sequence $(\mathbf{x}_k)_{k=0}^\infty$ that converges to some $\mathbf{x}$, since $\mathbb{R}^n$ is complete. However, $\mathbf{x} \in G_\alpha$ for some α , and since G_α is open, it contains an open ball $B(\mathbf{x}, r)$ for some $r > 0$. There exists an $N > 0$ such that $2^{-N} \|\mathbf{b} - \mathbf{a}\|_2 < r$, and thus $I_N \subset B(\mathbf{x}, r) \subset G_\alpha$ for all $n \geq N$. This gives a finite subcover of all these I_n , which is a contradiction. Thus, $[\mathbf{a}, \mathbf{b}]$ is compact.

Now let E be any closed and bounded subset of $\mathbb{R}^n$. Because E is bounded, it is contained in some n -cell $[\mathbf{a}, \mathbf{b}]$. Since E is closed and $[\mathbf{a}, \mathbf{b}]$ is compact, Proposition 5.5.2 guarantees that E is also compact. $\square$

Example 5.5.5.* The Heine-Borel theorem does not hold in general (infinite-dimensional) spaces. Consider the vector space $\mathbb{F}^\infty \subset \ell^2$ (see Example 1.1.6(iv)) defined to be the set of all infinite sequences $(x_1, x_2, \dots)$ with at most a finite number of nonzero entries.

The vector space $\mathbb{F}^\infty$ has an infinite basis of the form

$$B = \{(1, 0, 0, \dots), (0, 1, 0, \dots), \dots, (0, \dots, 0, 1, 0, \dots), \dots\}.$$

The unit sphere in this space (with the Euclidean metric) is closed and bounded, but not compact. Indeed, Example 5.1.13 and Exercise 5.7 give an example of an open cover of the unit ball that has no finite subcover.

5.5.2 Continuity and Compactness

Compactness is especially useful when coupled with continuity. In this section we prove that continuous functions preserve compact sets, that continuous functions on compact sets attain their supremum and infimum, and that continuous functions on compact sets are uniformly continuous.

Proposition 5.5.6. *The continuous image of a compact set is compact; that is, if $f : X \rightarrow Y$ is continuous and $K \subset X$ is compact, then $f(K) \subset Y$ is compact.*

Proof. If $(G_\alpha)_{\alpha \in J}$ is an open cover of $f(K)$, then $(f^{-1}(G_\alpha))_{\alpha \in J}$ is an open cover of K . Since K is compact, there is a finite subcover $(f^{-1}(G_{\alpha_k}))_{k=1}^n$. Hence, $(G_{\alpha_k})_{k=1}^n$ is a finite subcollection of $(G_\alpha)_{\alpha \in J}$ that covers $f(K)$, and thus $f(K)$ is compact. $\square$

Corollary 5.5.7 (Extreme Value Theorem). *If $f : X \rightarrow \mathbb{R}$ is continuous and $K \subset X$ is a nonempty compact set, then $f(K)$ contains its infimum and supremum.*

Proof. The image $f(K)$ is compact, hence closed and bounded in $\mathbb{R}$. Because it is bounded, its supremum and infimum both exist. Let M be the supremum, and let $(\mathbf{x}_n)_{n=0}^\infty$ be a sequence such that $f(\mathbf{x}_n) \rightarrow M$. Since K is compact, there is a subsequence $(\mathbf{x}_{n_i})_{i=0}^\infty$ converging to some value $\mathbf{x} \in K$. Since it is a subsequence, we must have $f(\mathbf{x}_{n_i}) \rightarrow M$, but continuity of f implies that $f(\mathbf{x}_{n_i}) \rightarrow f(\mathbf{x})$; see Theorem 5.3.19. Therefore, $M = f(\mathbf{x}) \in f(K)$. A similar argument shows that the infimum lies in $f(K)$. $\square$

Example 5.5.8.  Recall from Example 3.5.4 and Figures 3.5 and 3.6 that different metrics on a given space define different open balls. For example, the 1-norm unit sphere $S_1 = \{\mathbf{x} \in \mathbb{F}^n \mid \|\mathbf{x}\|_1 = 1\}$ is really a square when $n = 2$ and an octahedron when $n = 3$, whereas the 2-norm unit sphere is really a circle when $n = 2$ and a sphere when $n = 3$.

The 1-norm unit sphere S_1 is both closed and bounded (with respect to both the 1-norm and the 2-norm), so it is compact. Consequently, for any continuous function $f : \mathbb{F}^n \rightarrow \mathbb{R}$, the image $f(S_1)$ contains both its maximum and minimum.

In particular, if $\|\cdot\|$ is any norm on $\mathbb{F}^n$, then by Proposition 5.2.7 the map $f(\mathbf{x}) = \|\mathbf{x}\|$ is continuous (with respect to both the 1-norm and the 2-norm),

and hence there exists $\mathbf{x}_{max}, \mathbf{x}_{min} \in S_1$ such that $\sup_{\mathbf{x} \in S_1} \|\mathbf{x}\| = \|\mathbf{x}_{max}\|$ and $\inf_{\mathbf{x} \in S_1} \|\mathbf{x}\| = \|\mathbf{x}_{min}\|$.

This example plays an important role in the proof of the remarkable Theorem 5.8.7, which states that the open sets defined by any norm on a finite-dimensional vector space are the same as the open sets defined by any other norm on the same space.

Theorem 5.5.9. *If $K \subset X$ is compact and $f : K \rightarrow Y$ is continuous, then f is uniformly continuous on K .*

Proof. Let $\varepsilon > 0$ be given. For each $\mathbf{x} \in K$, there exists $\delta_{\mathbf{x}} > 0$ such that $\rho(f(\mathbf{x}), f(\mathbf{y})) < \varepsilon/2$ whenever $d(\mathbf{x}, \mathbf{y}) < \delta_{\mathbf{x}}$. Let $(B(\mathbf{x}, \frac{\delta_{\mathbf{x}}}{2}))_{\mathbf{x} \in K}$ be an open cover of K . Since K is compact, there exists a finite subcover $(B(\mathbf{x}_k, \frac{\delta_{\mathbf{x}_k}}{2}))_{k=1}^n$. So, given any $\mathbf{y} \in K$, there exists $k \in \{1, 2, \dots, n\}$ such that $d(\mathbf{y}, \mathbf{x}_k) < \frac{1}{2}\delta_{\mathbf{x}_k}$.

Let $\delta = \min\{\delta_{\mathbf{x}_k}\}_{k=1}^n$. If $\mathbf{y}, \mathbf{z} \in K$ with $d(\mathbf{y}, \mathbf{z}) < \delta/2$, then we have

$$d(\mathbf{x}_k, \mathbf{z}) \leq d(\mathbf{x}_k, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) < \frac{\delta_{\mathbf{x}_k}}{2} + \frac{\delta}{2} \leq \delta_{\mathbf{x}_k}.$$

Hence, $\rho(f(\mathbf{x}_k), f(\mathbf{z})) < \frac{\varepsilon}{2}$, and so

$$\rho(f(\mathbf{y}), f(\mathbf{z})) \leq \rho(f(\mathbf{y}), f(\mathbf{x}_k)) + \rho(f(\mathbf{x}_k), f(\mathbf{z})) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus, f is uniformly continuous. $\square$

5.5.3 Characterizations of Compactness

The definition of compactness is not always so easy to verify directly, but the Heine–Borel theorem gives a nice way to check that finite-dimensional spaces are compact. In this section we give several more useful characterizations of compactness and generalize the Heine–Borel theorem to arbitrary metric spaces.

Definition 5.5.10.

- (i) A collection $\mathcal{C}$ of sets in X has the finite intersection property if every finite subcollection of $\mathcal{C}$ has a nonempty intersection.
- (ii) The space X is sequentially compact if every sequence $(\mathbf{x}_k)_{k=0}^{\infty} \subset X$ has a convergent subsequence.
- (iii) The space X is totally bounded if for all $\varepsilon > 0$ the cover $\mathcal{C} = (B(\mathbf{x}, \varepsilon))_{\mathbf{x} \in X}$ has a finite subcover.
- (iv) A real number ε_0 is a Lebesgue number of an open cover $(G_{\alpha})_{\alpha \in J}$ if for all $\mathbf{x} \in X$ and for all $\varepsilon < \varepsilon_0$ the ball $B(\mathbf{x}, \varepsilon)$ is contained in some G_{α} .

Theorem 5.5.11. *Let (X, d) be a metric space. The following are equivalent:*

- (i) X is compact.

- (ii) Every collection $\mathcal{C}$ of closed sets in X with the finite intersection property has a nonempty intersection.
- (iii) X is sequentially compact.
- (iv) X is totally bounded and every open cover has a positive Lebesgue number (which depends on the cover).

Proof.

- (i) $\implies$ (ii) Assume X is compact. Let $\mathcal{C} = (F_\alpha)_{\alpha \in J}$ be a collection of closed sets with the finite intersection property. If $\bigcap_{\alpha \in J} F_\alpha = \emptyset$, then $(F_\alpha^c)_{\alpha \in J}$ is an open cover of X . Hence, there exists a finite subcover $(F_{\alpha_k}^c)_{k=1}^n$. But this implies that $\bigcap_{k=1}^n F_{\alpha_k} = \emptyset$, which is a contradiction.
- (ii) $\implies$ (iii) Let $(\mathbf{x}_k)_{k=0}^\infty \subset X$. For each $n \in \mathbb{N}$ let $B_n = \{\mathbf{x}_n, \mathbf{x}_{n+1}, \dots\}$. Thus $(\overline{B_n})_{n=0}^\infty$ is a collection of closed sets with the finite intersection property. By (ii), there exists $\mathbf{x} \in \bigcap_{k=1}^\infty \overline{B_k}$. For each k , choose $\mathbf{x}_{n_k}$ so that $n_k > n_{k-1}$ and $\mathbf{x}_{n_k} \in B(\mathbf{x}, 1/k)$. Thus, $\mathbf{x}_{n_k} \rightarrow \mathbf{x}$.
- (iii) $\implies$ (iv) We start by showing that every cover has a positive Lebesgue number. Assume $(G_\alpha)_{\alpha \in J}$ is an open cover of X . For each $\mathbf{x} \in X$ define

$$\varepsilon(\mathbf{x}) = \frac{1}{2} \sup\{\delta > 0 \mid \exists \alpha \in J \text{ with } B(\mathbf{x}, \delta) \subset G_\alpha\}.$$

Since $\mathbf{x}$ is an interior point of G_α , we have $\varepsilon(\mathbf{x}) > 0$ for all $\mathbf{x} \in X$. Define $\varepsilon_* = \inf_{\mathbf{x} \in X} \varepsilon(\mathbf{x})$. This is clearly a Lebesgue number of the cover, and we must show that $\varepsilon_* > 0$.

Since ε_* is an infimum, there exists a sequence $(\mathbf{x}_k)_{k=0}^\infty \subset X$ so that $\varepsilon(\mathbf{x}_k) \rightarrow \varepsilon_*$. Because X is sequentially compact, there is a convergent subsequence. Replacing the original sequence with the subsequence, we may assume³⁵ that $\mathbf{x}_k \rightarrow \mathbf{x}$ for some $\mathbf{x}$.

Let G_β be a member of the open cover such that $B(\mathbf{x}, \varepsilon(\mathbf{x})) \subset G_\beta$. Since $\mathbf{x}_k \rightarrow \mathbf{x}$, there exists $N \geq 0$ such that $d(\mathbf{x}, \mathbf{x}_n) < \frac{\varepsilon(\mathbf{x})}{2}$ whenever $n \geq N$. If $\mathbf{z} \in B(\mathbf{x}_n, \frac{\varepsilon(\mathbf{x})}{2})$, then $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{x}_n) + d(\mathbf{x}_n, \mathbf{z}) < \frac{\varepsilon(\mathbf{x})}{2} + \frac{\varepsilon(\mathbf{x})}{2} = \varepsilon(\mathbf{x})$, which implies that $\mathbf{z} \in B(\mathbf{x}, \varepsilon(\mathbf{x})) \subset G_\beta$. Thus, $B(\mathbf{x}_n, \frac{\varepsilon(\mathbf{x})}{2}) \subset G_\beta$, and $\varepsilon(\mathbf{x}_n) \geq \frac{\varepsilon(\mathbf{x})}{4} > 0$ for all $n \geq N$. This implies that $\varepsilon_* \geq \frac{\varepsilon(\mathbf{x})}{4} > 0$.

Now suppose X is not totally bounded. There is an $\varepsilon > 0$ such that the cover $(B(\mathbf{x}, \varepsilon))_{\mathbf{x} \in X}$ has no finite subcover. We define a sequence $(\mathbf{x}_k)_{k=0}^\infty$ as follows: choose any $\mathbf{x}_0$; since $B(\mathbf{x}_0, \varepsilon)$ does not cover X , there must be an $\mathbf{x}_1 \in B(\mathbf{x}_0, \varepsilon)^c$. Since the finite collection $(B(\mathbf{x}_0, \varepsilon), B(\mathbf{x}_1, \varepsilon))$ does not cover, there must be an $\mathbf{x}_2 \in X$ that is not in the union of these two balls. Continuing in this manner, we construct a sequence $(\mathbf{x}_k)_{k=0}^\infty$ so that $d(\mathbf{x}_k, \mathbf{x}_l) \geq \varepsilon$ for all $k \neq l$. Thus, $(\mathbf{x}_k)_{k=0}^\infty$ has no convergent subsequence, and X is not sequentially compact, which is a contradiction.

³⁵Since $\varepsilon(\mathbf{x})$ is not known to be continuous, we can't assume $\varepsilon(\mathbf{x}_n) \rightarrow \varepsilon(\mathbf{x})$.

- (iv) $\implies$ (i) Let $\mathcal{C} = (G_\alpha)_{\alpha \in J}$ be an open cover of X with Lebesgue number ε_* . Since X is totally bounded, it can be covered by a finite collection of balls $(B(\mathbf{x}_k, \varepsilon))_{k=1}^n$ where $\varepsilon < \varepsilon_*$. Each ball $B(\mathbf{x}_k, \varepsilon)$ is contained in some G_{α_k} of the open cover, so the finite collection $(G_{\alpha_k})_{k=1}^n$ is a subcover. Thus, X is compact. $\square$

Theorem 5.5.12 (Generalized Heine–Borel). *A metric space X is compact if and only if it is complete and totally bounded.*

Proof. ($\implies$) Assume that X is compact. By the previous theorem, it is totally bounded. If $(\mathbf{x}_k)_{k=0}^\infty$ is a Cauchy sequence in X , then by the previous theorem, it has a convergent subsequence. By Proposition 5.4.9 any Cauchy sequence with a convergent subsequence must converge. Thus, $(\mathbf{x}_k)_{k=0}^\infty$ converges. Since this holds for every Cauchy sequence, X is complete.

($\impliedby$) Assume that X is complete and totally bounded. It suffices to show that each sequence $(\mathbf{x}_k)_{k=0}^\infty \subset X$ has a convergent subsequence. Let $E = \{\mathbf{x}_k\}_{k=0}^\infty$ be the set of points of the sequence. If E is finite, it clearly has a convergent subsequence, since infinitely many of the $\mathbf{x}_k$ must be equal to the same point. Thus, we may assume that E is infinite.

Since X is totally bounded, we can cover it with a finite number of open balls of radius 1. One of these open balls contains infinitely many points of E . Denote this as B_1 . Now cover X with finitely many balls of radius $\frac{1}{2}$. We can pick one whose intersection with B_1 contains infinitely many points of E . Call this B_2 . Repeating yields a sequence of open balls $(B_k)_{k=1}^\infty$ of radius $\frac{1}{k}$. For each ball choose $\mathbf{x}_{n_k} \in B_k$. The sequence $(\mathbf{x}_{n_k})_{k=0}^\infty$ is Cauchy, and it converges, since X is complete. Thus, the sequence $(\mathbf{x}_n)_{n=0}^\infty$ has a convergent subsequence $(\mathbf{x}_{n_k})$. $\square$

5.5.4 * Subspaces

If d is a metric on X , then d induces a metric on every $Y \subset X$; that is, restricting d to the set $Y \times Y$ defines a function $\rho : Y \times Y \rightarrow [0, \infty)$ that itself satisfies all the conditions for a metric on the space Y . We say that Y inherits the metric ρ from X .

Example 5.5.13. If $X = \mathbb{R}^2$ with the standard metric $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$, and if $Y = S^1 = \{(\cos(t), \sin(t)) \mid t \in [0, 2\pi)\}$ is the unit circle, then the induced metric on S^1 is just the usual distance between the points of S^1 , thought of as points in the plane. So if $\mathbf{a} = (\cos(t), \sin(t))$ and $\mathbf{b} = (\cos(s), \sin(s))$, then $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$. In particular, we have $d((\cos(0), \sin(0)), (\cos(s), \sin(s))) \rightarrow 0$ as $s \rightarrow 2\pi$.

Definition 5.5.14. For each $\mathbf{x} \in Y \subset X$ and $r > 0$, we write $B_X(\mathbf{x}, r) = \{\mathbf{y} \in X \mid d(\mathbf{x}, \mathbf{y}) < r\}$ and $B_Y(\mathbf{x}, r) = \{\mathbf{y} \in Y \mid \rho(\mathbf{x}, \mathbf{y}) < r\} = B_X(\mathbf{x}, r) \cap Y$.

A set E is open in Y or is open relative to Y if E is an open set in the metric space (Y, ρ) ; that is, E is open in Y if for every $\mathbf{x} \in E$ there is some open ball $B_Y(\mathbf{x}, r) \subset E$.

Remark 5.5.15. Theorem 5.1.12 applied to the space Y shows that the ball $B_Y(\mathbf{x}, r)$ is open in Y . But it is not necessarily open in X , as we see in the next example.

Example 5.5.16. If $Y = [0, 1] \subset \mathbb{R}$, then

$$B_Y(0, 1/2) = \{y \in [0, 1] \mid |y - 0| < 1/2\} = [0, 1/2)$$

is open in $[0, 1]$ but not open in $\mathbb{R}$.

The subspace $[0, 1]$ with its induced metric (from the usual metric on $\mathbb{R}$) has open sets that look like $[0, x)$, (x, y) , and $(x, 1]$ and unions of such sets.

Any set that is open in Y is the intersection of Y with an open set in X , as the next proposition shows.

Proposition 5.5.17. *Let (X, d) be a metric space, and let $Y \subset X$ be a subset with the induced metric ρ , so that (Y, ρ) is itself a metric space. A subset E of (Y, ρ) is open if and only if there is an open subset $U \subset X$ such that $E = U \cap Y$. Similarly, a subset F of (Y, ρ) is closed if and only if there is a closed set $C \subset X$ such that $F = C \cap Y$.*

Proof. If E is an open set in (Y, ρ) , then around every point $\mathbf{x} \in E$, there is a ball $B_Y(\mathbf{x}, r_{\mathbf{x}}) \subset E$. Let $U = \bigcup_{\mathbf{x} \in E} B_X(\mathbf{x}, r_{\mathbf{x}}) \subset X$. Since U is the union of open balls in X , it is open in X . For each $\mathbf{x} \in E$ we certainly have $\mathbf{x} \in U$, thus $E \subset U \cap Y$. But we also have $B_X(\mathbf{x}, r_{\mathbf{x}}) \cap Y = B_Y(\mathbf{x}, r_{\mathbf{x}}) \subset E$, so $U \cap Y \subset E$.

Conversely, if U is open in X , then for any $\mathbf{x} \in U \cap Y$, we have some ball $B_X(\mathbf{x}, r_{\mathbf{x}}) \subset U$. Therefore, $B_Y(\mathbf{x}, r_{\mathbf{x}}) = B_X(\mathbf{x}, r_{\mathbf{x}}) \cap Y \subset U \cap Y$, so $U \cap Y$ is open in (Y, ρ) .

The statement about closed sets follows from taking the complement of the open sets. $\square$

Proposition 5.5.18. *Assume that Y is a subspace of X . If $E \subset Y$, then the closure of E in the subspace (Y, ρ) is given by $\overline{E} \cap Y$, where $\overline{E}$ is the closure of E in (X, d) .*

Proof. Let $\tilde{E}$ denote the closure of E in Y with respect to ρ . Clearly $\overline{E} \cap Y$ is closed in Y and contains E , so $\tilde{E}$ is contained in $\overline{E} \cap Y$. Conversely, $\tilde{E}$ is closed in Y , and thus, by Proposition 5.5.17, there must be a set $F \subset X$ that is closed with respect to d such that $F \cap Y = \tilde{E}$. But this implies that $E \subset F$; hence, $\overline{E} \subset F$, and thus $\overline{E} \cap Y \subset F \cap Y = \tilde{E}$. $\square$

Example 5.5.19. The integers $\mathbb{Z}$ as a subspace of $\mathbb{R}$, with the usual metric, has the property that every element is an open set. Thus, the family $\mathcal{T}$ of all open sets is the power set of $\mathbb{Z}$, sometimes denoted $2^{\mathbb{Z}}$. This is the most open

sets possible. As a result, every function with domain $\mathbb{Z}$ is continuous (but continuous functions from $\mathbb{F}^n$ to $\mathbb{Z}$ must be constant). We note that these are the same open sets as those we got with the discrete metric on $\mathbb{Z}$.

Now that we have defined an induced metric on subsets, there are actually two different ways to define compactness of any subset $Y \subset X$. The first is as a subset of X with the usual metric on X (that is, in terms of open sets of X). The second is as a subspace, with the induced metric (that is, in terms of open sets of Y). Fortunately, these are equivalent.

Proposition 5.5.20. *A subset $Y \subset X$ is compact with respect to open sets of X if and only if it is compact with respect to open sets of Y , as a subspace with the induced metric.*

Proof. The proof is Exercise 5.34. $\square$

5.6 Uniform Convergence and Banach Spaces

5.6.1 Pointwise and Uniform Convergence

There are at least two distinct notions of convergence that are useful when considering a sequence $(f_n)_{n=0}^\infty$ of functions: *pointwise convergence* and *uniform convergence*. Pointwise convergence is convergence of the sequences $(f_n(x))_{n=0}^\infty$ for each x in the domain, whereas uniform convergence is convergence in the space of functions with respect to the L^∞ -norm.

Definition 5.6.1. *For any sequence of functions $(f_n)_{n=0}^\infty$ from a set X into a metric space Y , we can evaluate all the functions at a single point x of the domain, which gives a sequence $(f_n(x))_{n=0}^\infty \subset Y$. If for every choice of $x \in X$, the sequence $(f_n(x))_{n=0}^\infty$ converges in Y , then we can define a new function f by setting $f(x) = \lim_{n \rightarrow \infty} f_n(x)$. In this case we say that the sequence converges pointwise, or that f is the pointwise limit of (f_n) .*

Example 5.6.2. The sequence $(f_n)_{n=0}^\infty$ of functions on $(0, 1)$ defined by $f_n(x) = x^n$ converges pointwise to the zero function, since for each $x \in (0, 1)$ we have $x^n \rightarrow 0$ as $n \rightarrow \infty$.

Pointwise convergence is often not strong enough to be very useful. A much stronger notion is that of *uniform convergence*, which is convergence in the L^∞ -norm.

Definition 5.6.3. *Let $(f_k)_{k=0}^\infty$ be a sequence of bounded functions from a set X into a normed space Y . If $(f_k)_{k=0}^\infty$ converges to f in the L^∞ -norm, then we say that the sequence $(f_k)_{k=0}^\infty$ converges uniformly to f .*

Unexample 5.6.4. Although the sequence $(x^n)_{n=0}^\infty$ of functions on $(0, 1)$ converges pointwise to 0, it does not converge uniformly to 0 on $(0, 1)$. To see this, observe that

$$\|f_n(x) - 0\|_{L^\infty} = \sup_{x \in (0,1)} |x^n| = 1$$

for all n . In the next proposition we see that if there were a uniform limit, it would have to be the same as the pointwise limit.

Proposition 5.6.5. *Uniform convergence implies pointwise convergence. That is to say, if $(f_n)_{n=0}^\infty$ is a uniformly convergent sequence with limit f , then f_n converges pointwise to f .*

Proof. The proof is Exercise 5.35. $\square$

5.6.2 Banach Spaces

Completeness is a very useful property of a metric space. Normed vector spaces that are complete are especially useful.

Definition 5.6.6. *A complete normed linear space is called a Banach space.*

Example 5.6.7. Most of the normed linear spaces that we work with in applied mathematics are Banach spaces. We showed in Theorem 5.4.16 that $(\mathbb{F}^n, \|\cdot\|_p)$ is a Banach space. Below we show that $(C([a, b]; \mathbb{R}), \|\cdot\|_{L^\infty})$ is also a Banach space. Additional examples include the spaces ℓ^p for $1 \leq p \leq \infty$ (see Example 1.1.6(vi)).

Theorem 5.6.8. *The space $(C([a, b]; \mathbb{F}), \|\cdot\|_{L^\infty})$ is a Banach space.*

Proof. We need only prove that $(C([a, b]; \mathbb{F}), \|\cdot\|_{L^\infty})$ is complete. We do this by showing (i) any Cauchy sequence converges pointwise, which defines a candidate function for the L^∞ limit; (ii) the convergence is actually uniform, so the candidate really is the L^∞ limit (but might not lie in the space $C([a, b]; \mathbb{F})$); and (iii) the limit does indeed lie in $C([a, b]; \mathbb{F})$.

- (i) Let $(f_k)_{k=0}^\infty \subset C([a, b]; \mathbb{F})$ be a Cauchy sequence. For a given $\varepsilon > 0$, there exists $N > 0$ such that $\|f_m - f_n\|_{L^\infty} < \varepsilon$ whenever $m, n \geq N$. For a fixed $x \in [a, b]$ we have that $|f_m(x) - f_n(x)| \leq \sup_{t \in [a, b]} |f_m(t) - f_n(t)| < \varepsilon$, and thus the sequence $(f_k(x))_{k=0}^\infty$ is a Cauchy sequence. Since $\mathbb{F}$ is complete, the sequence converges. Define f as the pointwise limit of $(f_n)_{n=0}^\infty$; that is, for each $x \in [a, b]$ define $f(x) = \lim_{k \rightarrow \infty} f_k(x)$.
- (ii) We show that $(f_k)_{k=0}^\infty$ converges to f uniformly. Given $\varepsilon > 0$, there exists $N > 0$ such that $\|f_n - f_m\|_{L^\infty} < \varepsilon/2$ whenever $m, n \geq N$. By Corollary 5.3.22 we can pass the limit through pointwise; that is, for each $x \in [a, b]$ we have

$$|f(x) - f_m(x)| = \lim_{n \rightarrow \infty} |f_n(x) - f_m(x)| \leq \varepsilon/2 < \varepsilon.$$

Thus, it follows that $\|f - f_m\|_{L^\infty} < \varepsilon$.

- (iii) It remains to prove that f is continuous on $[a, b]$. For each $\varepsilon > 0$, choose N such that $\|f_m - f_n\|_{L^\infty} < \frac{\varepsilon}{3}$ whenever $m, n \geq N$. By taking the limit as $n \rightarrow \infty$, we have that $\|f_m - f\|_{L^\infty} \leq \frac{\varepsilon}{3}$ whenever $m \geq N$. Theorem 5.5.9 states that any continuous function on a compact set is uniformly continuous. Therefore, for any $m \geq N$, we know f_m is uniformly continuous on $[a, b]$, and thus there is a $\delta > 0$ such that $|f_m(x) - f_m(y)| < \frac{\varepsilon}{3}$ whenever $|x - y| < \delta$. Thus, we have that

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_m(x)| + |f_m(x) - f_m(y)| + |f_m(y) - f(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Therefore, f is continuous on $[a, b]$, which completes the proof. $\square$

The next corollary is an immediate consequence of Theorem 5.6.8.

Corollary 5.6.9. *If a sequence of continuous functions converges uniformly to a function f , then f is also continuous.*

5.6.3 Sums in a Banach Space

Throughout the remainder of this section, assume that $(X, \|\cdot\|)$ is a Banach space.

Definition 5.6.10. *Consider a sequence $(\mathbf{x}_k)_{k=0}^\infty \subset X$. We say that the series $\sum_{k=0}^\infty \mathbf{x}_k$ converges in X if the sequence $(\mathbf{s}_k)_{k=0}^\infty$ of partial sums, defined by $\mathbf{s}_n = \sum_{k=0}^n \mathbf{x}_k$, converges in X ; otherwise, we say that the series diverges.*

Definition 5.6.11. *Assume $(\mathbf{x}_k)_{k=0}^\infty$ is a sequence in X . The series $\sum_{k=0}^\infty \mathbf{x}_k$ is said to converge absolutely if the series $\sum_{k=0}^\infty \|\mathbf{x}_k\|$ converges in $\mathbb{R}$.*

Remark 5.6.12. If a series converges absolutely in X , then for every $\varepsilon > 0$ there is an N such that the sum $\sum_{k=n}^\infty \|\mathbf{x}_k\| < \varepsilon$ whenever $n \geq N$.

Proposition 5.6.13. *Let $(\mathbf{x}_k)_{k=0}^\infty$ be a sequence in X . If the series $\sum_{k=0}^\infty \mathbf{x}_k$ converges absolutely, then it converges in X .*

Proof. It suffices to show that the sequence of partial sums $(\mathbf{s}_k)_{k=0}^\infty$ converges in X . Let $\varepsilon > 0$ be given. If the series $\sum_{k=0}^\infty \mathbf{x}_k$ converges absolutely, then there exists N such that $\sum_{k=n}^\infty \|\mathbf{x}_k\| < \varepsilon$ whenever $n \geq N$. Thus, we have

$$\|\mathbf{s}_n - \mathbf{s}_m\| = \left\| \sum_{k=m+1}^n \mathbf{x}_k \right\| \leq \sum_{k=m+1}^n \|\mathbf{x}_k\| \leq \sum_{k=m+1}^\infty \|\mathbf{x}_k\| < \varepsilon$$

whenever $n > m \geq N$. This implies that the sequence of partial sums is Cauchy, and hence it converges, since X is complete. $\square$

Example 5.6.14. For each $n \in \mathbb{N}$ let $f_n \in C([0, 2]; \mathbb{R})$ be the function $f_n(x) = x^n/n!$. The sum $\sum_{n=0}^{\infty} f_n$ converges absolutely because the series

$$\sum_{n=0}^{\infty} \|f_n\|_{L^\infty} = \sum_{n=0}^{\infty} \sup_{x \in [0, 2]} |x^n|/n! = \sum_{n=0}^{\infty} 2^n/n! = e^2$$

converges in $\mathbb{R}$.

Theorem 5.6.15. If a sum $\sum_{k=0}^{\infty} \mathbf{x}_k$ converges absolutely to $\mathbf{x} \in X$, then for any rearrangement of the terms, the rearranged sum converges absolutely to $\mathbf{x}$. That is, if $f: \mathbb{N} \rightarrow \mathbb{N}$ is a bijection, then $\sum_{k=0}^{\infty} \mathbf{x}_{f(k)}$ also converges absolutely to $\mathbf{x}$.

Proof. For any $\varepsilon > 0$ choose $N > 0$ such that $\sum_{n=N}^{\infty} \|\mathbf{x}_n\| < \varepsilon/2$. Since N is finite, there must be some $M \geq N$ so that the set $\{1, 2, \dots, N\}$ is a subset of $\{f(1), \dots, f(M)\}$. For any $n > M$ let $E_n = \{f(1), \dots, f(n)\} \setminus \{1, 2, \dots, N\} \subset \{N+1, N+2, \dots\}$. We have

$$\begin{aligned} \left\| \mathbf{x} - \sum_{k=0}^n \mathbf{x}_{f(k)} \right\| &= \left\| \mathbf{x} - \sum_{k=0}^N \mathbf{x}_k + \sum_{k=0}^N \mathbf{x}_k - \sum_{k=0}^n \mathbf{x}_{f(k)} \right\| \\ &\leq \left\| \mathbf{x} - \sum_{k=0}^N \mathbf{x}_k \right\| + \left\| \sum_{k=0}^N \mathbf{x}_k - \sum_{k=0}^n \mathbf{x}_{f(k)} \right\| \\ &< \frac{\varepsilon}{2} + \sum_{k \in E_n} \|\mathbf{x}_k\| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Hence, the rearranged series converges to $\mathbf{x}$.

The fact that the rearranged sum converges absolutely follows from applying the same argument to the series $\sum_{k=0}^{\infty} \|\mathbf{x}_k\|$. $\square$

Remark 5.6.16. The converse to the previous theorem is false. In any infinite-dimensional Banach space there are series that converge, regardless of the rearrangement, and yet are not absolutely convergent (see [DR50]).

5.7 The Continuous Linear Extension Theorem

In this section we prove several important theorems about bounded functions and Banach spaces. We first prove that the space of bounded linear transformations into a Banach space is itself a Banach space. We then prove that the space of bounded functions into a Banach space is also a Banach space. This means we can use all the tools we have developed about convergence in Banach spaces when studying bounded linear transformations or bounded functions.

The main result of this section is a powerful theorem known as the *continuous linear extension theorem*, also known as the *bounded linear transformation (BLT) theorem*. This result tells us that to define a bounded linear transformation $T: Z \rightarrow X$ from a normed linear space Z to a Banach space X , it suffices to define

T on a dense subspace of Z . This is very useful because it is often easy to define a transformation having the properties we want on some dense subspace, while the definition on the entire space might be more difficult.

An important application of this theorem is the construction of the integral of a Banach-valued function. We do this for single-variable functions in Section 5.10 and for multivariable functions in Section 8.1. We also use it later in Chapter 8 to define the Lebesgue integral. In every case, the idea is simple: first define the integral on some of the simplest functions imaginable—step functions—and then use the continuous linear extension theorem to extend to more general functions.

5.7.1 Bounded Linear Transformations

The convergence results and other tools developed previously in this chapter are useful in many settings. The following result tells us we can also use them when studying matrices and other bounded operators (see Definition 3.5.10).

Theorem 5.7.1. *Let $(X, \|\cdot\|_X)$ be a normed linear space, and let $(Y, \|\cdot\|_Y)$ be a Banach space. The space $\mathcal{B}(X; Y)$ of bounded linear transformations from X into Y is a Banach space when endowed with the induced norm $\|\cdot\|_{X,Y}$. In particular, the space $X^* = \mathcal{B}(X; \mathbb{F})$ of bounded linear functionals is a Banach space.*

Proof. The space $\mathcal{B}(X; Y)$ is a normed linear space by Theorem 3.5.11, so we need only prove it is complete. Let $(A_k)_{k=0}^\infty \subset \mathcal{B}(X; Y)$ be a Cauchy sequence. We prove that it converges in three steps: (i) construct a function $A : X \rightarrow Y$ that is a candidate for the limit, (ii) show that A is in $\mathcal{B}(X; Y)$, and (iii) show that $A_k \rightarrow A$ in the norm $\|\cdot\|_{X,Y}$.

- (i) Define $A : X \rightarrow Y$ as follows. Given any nonzero $\mathbf{x} \in X$ and $\varepsilon > 0$, there exists $N > 0$ such that $\|A_m - A_n\|_{X,Y} < \varepsilon / \|\mathbf{x}\|_X$, whenever $m, n \geq N$. Define $\mathbf{y}_k = A_k \mathbf{x}$ for each $k \in \mathbb{N}$. Thus,

$$\|\mathbf{y}_m - \mathbf{y}_n\|_Y = \|A_m \mathbf{x} - A_n \mathbf{x}\|_Y \leq \|A_m - A_n\|_{X,Y} \|\mathbf{x}\|_X < \varepsilon$$

whenever $m, n \geq N$. Thus, $(\mathbf{y}_k)_{k=0}^\infty$ is a Cauchy sequence in the Banach space Y , and so it must converge to some $\mathbf{y} \in Y$. This procedure defines a mapping from $X \setminus \mathbf{0}$ to Y , and we can extend it to all of X by sending $\mathbf{0}$ to $\mathbf{0}$. Call the resulting map A .

- (ii) We now show that $A \in \mathcal{B}(X; Y)$. First, we see that A is linear since

$$A(a\mathbf{x}_1 + b\mathbf{x}_2) = \lim_{k \rightarrow \infty} A_k(a\mathbf{x}_1 + b\mathbf{x}_2) = a \lim_{k \rightarrow \infty} A_k \mathbf{x}_1 + b \lim_{k \rightarrow \infty} A_k \mathbf{x}_2 = aA\mathbf{x}_1 + bA\mathbf{x}_2.$$

To show that A is a bounded linear transformation, again fix some $\varepsilon > 0$ and choose an N such that $\|A_m - A_n\|_{X,Y} < \varepsilon$ when $m, n \geq N$. Thus,

$$\|A_m \mathbf{x} - A_n \mathbf{x}\|_Y \leq \|A_m - A_n\|_{X,Y} \|\mathbf{x}\|_X < \varepsilon \|\mathbf{x}\|_X.$$

Taking the limit as $m \rightarrow \infty$, we have that

$$\|A\mathbf{x} - A_n \mathbf{x}\|_Y \leq \varepsilon \|\mathbf{x}\|_X. \quad (5.7)$$

It follows that

$$\|A\mathbf{x}\|_Y \leq \|A\mathbf{x} - A_n\mathbf{x}\|_Y + \|A_n\mathbf{x}\|_Y \leq \varepsilon\|\mathbf{x}\|_X + \|A_n\|_{X,Y}\|\mathbf{x}\|_X.$$

By Proposition 5.4.7, the sequence $(A_k)_{k=0}^\infty$ is bounded; hence, there exists $M > 0$ such that $\|A_n\|_{X,Y} \leq M$ for each $n \in \mathbb{N}$. Thus, $\|A\mathbf{x}\|_Y \leq (\varepsilon + M)\|\mathbf{x}\|_X$, and

$$\|A\|_{X,Y} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_Y}{\|\mathbf{x}\|_X} \leq (\varepsilon + M).$$

Therefore, A is in $\mathcal{B}(X; Y)$.

- (iii) We conclude the proof by showing that $A_k \rightarrow A$ with respect to the norm $\|\cdot\|_{X,Y}$. By (5.7) we have $\frac{\|A\mathbf{x} - A_n\mathbf{x}\|_Y}{\|\mathbf{x}\|_X} \leq \varepsilon$ whenever $n \geq N$ and $\mathbf{x} \neq \mathbf{0}$. By taking the supremum over all $\mathbf{x} \neq \mathbf{0}$, we have that $\|A - A_n\|_{X,Y} \leq \varepsilon$, whenever $n \geq N$. $\square$

Example 5.7.2. Let $(X, \|\cdot\|_X)$ be a Banach space. Since $\mathcal{B}(X)$ is also a Banach space, it makes sense to define infinite series in $\mathcal{B}(X)$, provided the series converge.

We define the exponential of a bounded operator $A \in \mathcal{B}(X)$ as $\exp(A) = \sum_{k=0}^\infty \frac{A^k}{k!}$, where $A^0 = I$. If $\|\cdot\|$ is the operator norm, then the series $\exp(A)$ converges absolutely if the series $\sum_{k=0}^\infty \frac{\|A^k\|}{k!}$ of real numbers converges. This is straightforward to check:

$$\sum_{k=0}^\infty \left\| \frac{A^k}{k!} \right\| \leq \sum_{k=0}^\infty \frac{\|A\|^k}{k!} = e^{\|A\|} < \infty.$$

Therefore, the operator $\exp(A)$ is defined for any bounded operator A .

Example 5.7.3. Let $(X, \|\cdot\|_X)$ be a Banach space. Let $A \in \mathcal{B}(X)$ be a bounded operator with $\|A\| < 1$. The *Neumann series* of A is the sum $\sum_{k=0}^\infty A^k$. It is the analogue of the geometric series. This is well defined, since the series is absolutely convergent. If $\|\cdot\|$ is the operator norm, then

$$\sum_{k=0}^\infty \|A^k\| \leq \sum_{k=0}^\infty \|A\|^k = \frac{1}{1 - \|A\|} < \infty.$$

Proposition 5.7.4. Let $(X, \|\cdot\|_X)$ be a Banach space. If $A \in \mathcal{B}(X)$ satisfies $\|A\| < 1$, then $I - A$ is invertible. Moreover, we have that $\sum_{k=0}^\infty A^k = (I - A)^{-1}$, and thus $\|(I - A)^{-1}\| \leq (1 - \|A\|)^{-1}$.

Proof. The proof is Exercise 5.43. $\square$

5.7.2 Bounded Functions

Recall from Proposition 3.5.9 that for any set S , the space $L^\infty(S; X)$ of all bounded functions from S into a normed linear space $(X, \|\cdot\|_X)$ with the sup norm $\|f\|_{L^\infty} = \sup_{t \in S} \|f(t)\|$ is a normed linear space. The next theorem shows it is also complete.

Theorem 5.7.5. *Let $(X, \|\cdot\|_X)$ be a Banach space. For any set S , the space $(L^\infty(S; X), \|\cdot\|_{L^\infty})$ is a Banach space.*

Proof. It suffices to show that the space is complete. Let $(f_k)_{k=0}^\infty \subset L^\infty(S; X)$ be a Cauchy sequence. For each fixed $t \in S$, the sequence $(f_k(t))_{k=0}^\infty$ is Cauchy and thus converges. Define $f(t) = \lim_{k \rightarrow \infty} f_k(t)$.

It suffices to show that $\|f - f_n\|_{L^\infty} \rightarrow 0$ as $n \rightarrow \infty$ and that $\|f\|_{L^\infty} < \infty$. As mentioned in Corollary 5.3.22, we can pass the limit through the norm. Specifically, given $\varepsilon > 0$, choose $N > 0$ such that $\|f_n - f_m\|_{L^\infty} < \varepsilon/2$ whenever $m, n \geq N$. Thus, for each $s \in S$, we have

$$\|f(s) - f_m(s)\| = \lim_{n \rightarrow \infty} \|f_n(s) - f_m(s)\| \leq \frac{\varepsilon}{2} < \varepsilon.$$

It follows that $\|f - f_m\|_{L^\infty} < \varepsilon$, which implies $\|f - f_n\|_{L^\infty} \rightarrow 0$.

To see that f is bounded, note that since $(f_k)_{k=0}^\infty$ is Cauchy, it is bounded (see Proposition 5.4.7), so there is some $M < \infty$ such that $\|f_n\|_{L^\infty} < M$. Again, passing the limit through the norm gives $\|f\|_{L^\infty} \leq M < \infty$. $\square$

5.7.3 The Continuous Linear Extension Theorem

The next theorem provides the key tool for defining integration. We use this theorem at the end of this chapter to define single-variable Banach-valued integrals, and again when we study Lebesgue integration in Chapter 8.

This theorem says that if we know how to define a continuous linear transformation on some dense subspace of a Banach space, then we can extend the transformation uniquely to a linear transformation on the whole space. This theorem is often known as the *bounded linear transformation theorem* because a linear transformation is bounded if and only if it is continuous,³⁶ and so the theorem says any bounded linear transformation on a dense subspace extends uniquely to a bounded linear transformation on the whole space.

Theorem 5.7.6 (Continuous Linear Extension Theorem). *Let $(Z, \|\cdot\|_Z)$ be a normed linear space, $(X, \|\cdot\|_X)$ a Banach space, and $S \subset Z$ a dense subspace of Z . If $T : S \rightarrow X$ is a bounded linear transformation, then T has a unique linear extension to $\bar{T} \in \mathcal{B}(Z; X)$ satisfying $\|\bar{T}\| = \|T\|$.*

Proof. For $z \in Z$, since S is dense, there exists a sequence $(s_k)_{k=0}^\infty$ in S that converges to z . Since it is a convergent sequence, it is Cauchy by Proposition 5.4.4. Moreover, since T is bounded on S , it is uniformly continuous there (see Proposition 5.4.22), and thus $(T(s_k))_{k=0}^\infty$ is also a Cauchy sequence by Theorem 5.4.24. Moreover, it converges, since X is a Banach space.

³⁶by Proposition 5.4.22

For $\mathbf{z} \in Z$, define $\bar{T}(\mathbf{z}) = \lim_{k \rightarrow \infty} T(\mathbf{s}_k)$. We now show this is well defined. Let $(\mathbf{s}'_k)_{k=0}^\infty$ be any other sequence in S converging to $\mathbf{z}$. For any $\varepsilon > 0$, by uniform continuity of T there is a $\delta > 0$ such that $\|T(a) - T(b)\|_X < \varepsilon/2$ whenever $\|a - b\|_Z < \delta$. Choose $K > 0$ such that $\|\mathbf{s}_k - \mathbf{z}\|_Z < \delta$ and $\|\mathbf{s}'_k - \mathbf{z}\|_Z < \delta$ whenever $k \geq K$. Thus, we have

$$\|T(\mathbf{s}'_k) - \bar{T}(\mathbf{z})\|_X \leq \|T(\mathbf{s}'_k) - T(\mathbf{s}_k)\|_X + \|T(\mathbf{s}_k) - \bar{T}(\mathbf{z})\|_X < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

whenever $k \geq K$. Therefore, $T(\mathbf{s}'_k) \rightarrow \bar{T}(\mathbf{z})$, and the value of $\bar{T}(\mathbf{z})$ is independent of the choice of $(\mathbf{s}_k)_{k=0}^\infty$.

It remains to prove that $\bar{T}$ is linear, that $\|\bar{T}\| = \|T\|$, and that $\bar{T}$ is unique. The linearity of $\bar{T}$ follows from the linearity of T . If $(\mathbf{s}_k)_{k=0}^\infty \subset S$ converges to $\mathbf{z} \in Z$ and $(\hat{\mathbf{s}}_k)_{k=0}^\infty \subset S$ converges to $\hat{\mathbf{z}} \in Z$, then for $a, b \in \mathbb{F}$ we have

$$\bar{T}(a\mathbf{z} + b\hat{\mathbf{z}}) - a\bar{T}(\mathbf{z}) - b\bar{T}(\hat{\mathbf{z}}) = \lim_{k \rightarrow \infty} (T(a\mathbf{s}_k + b\hat{\mathbf{s}}_k) - aT(\mathbf{s}_k) - bT(\hat{\mathbf{s}}_k)) = \mathbf{0}.$$

To see the norm of $\bar{T}$, note first that if $(\mathbf{s}_k)_{k=0}^\infty \subset S$ converges to $\mathbf{z} \in Z$, then

$$\|\bar{T}(\mathbf{z})\| = \left\| \lim_{k \rightarrow \infty} T(\mathbf{s}_k) \right\| = \lim_{k \rightarrow \infty} \|T(\mathbf{s}_k)\| \leq \lim_{k \rightarrow \infty} \|T\| \|\mathbf{s}_k\| = \|T\| \|\mathbf{z}\|.$$

Therefore, we have

$$\|\bar{T}\| = \sup_{\substack{\mathbf{z} \in Z \\ \mathbf{z} \neq \mathbf{0}}} \frac{\|\bar{T}(\mathbf{z})\|}{\|\mathbf{z}\|} \leq \|T\|,$$

but $\bar{T}(\mathbf{s}) = T(\mathbf{s})$ for all $\mathbf{s} \in S$, so

$$\|\bar{T}\| = \sup_{\substack{\mathbf{z} \in Z \\ \mathbf{z} \neq \mathbf{0}}} \frac{\|\bar{T}(\mathbf{z})\|}{\|\mathbf{z}\|} \geq \sup_{\substack{\mathbf{s} \in S \\ \mathbf{s} \neq \mathbf{0}}} \frac{\|T(\mathbf{s})\|}{\|\mathbf{s}\|} = \|T\|.$$

Finally, for uniqueness, suppose there were another extension $\hat{T}$ of T on Z . If $\hat{T}(\mathbf{z}) \neq \bar{T}(\mathbf{z})$ for some $\mathbf{z} \in Z \setminus S$, then for some sequence $(\mathbf{s}_k)_{k=0}^\infty \subset S$ that converges to $\mathbf{z}$, we have $\mathbf{0} = \lim_{k \rightarrow \infty} (\bar{T} - \hat{T})(\mathbf{s}_k) = \hat{T}(\mathbf{z}) - \bar{T}(\mathbf{z}) \neq \mathbf{0}$, which is a contradiction. Thus, the extension is unique. $\square$

5.7.4 * Invertible Operators

Proposition 5.7.4 gave a way to compute the inverse of $I - A$ for $\|A\| < 1$. We now need to study some basic properties of more general inverses. Recall that for any matrix the determinant is nonzero if and only if the inverse exists. Moreover, the function $\det : M_n(\mathbb{F}) \rightarrow \mathbb{F}$ is continuous, so the inverse image of 0 (the set $\det^{-1}(0) = \{A \in M_n(\mathbb{F}) \mid \det(A) = 0\}$) is closed, which means its complement, the set of all invertible matrices (often denoted $\text{GL}_n(\mathbb{F})$), is open.

Similarly, the adjugate of A is clearly a continuous function in the entries of A , so Cramer's rule (or rather Corollary 2.9.23) tells us that the inverse map $A \mapsto A^{-1}$ must be a continuous function of A .

The next proposition tells us that these two results hold for bounded linear operators on an arbitrary Banach space—not just for matrices. The difficulty in proving this comes from the fact that we have neither a determinant function nor an analogue of Cramer's rule.

Proposition 5.7.7. *Let $(X, \|\cdot\|)$ be a Banach space and define*

$$\mathrm{GL}(X) = \{A \in \mathcal{B}(X) \mid A^{-1} \in \mathcal{B}(X)\}.$$

The set $\mathrm{GL}(X)$ is open in $\mathcal{B}(X)$, and the function $\mathrm{Inv}(A) = A^{-1}$ is continuous on $\mathrm{GL}(X)$.

Proof. Let $A \in \mathrm{GL}(X)$. We claim that $B(A, r) \subset \mathrm{GL}(X)$ for $r = \|A^{-1}\|^{-1}$. This shows that $\mathrm{GL}(X)$ is open.

To see the claim, choose $L \in B(A, r)$ and write $L = A(I - A^{-1}(A - L))$. Since $\|A - L\| < r$, we have that

$$\|A^{-1}(A - L)\| \leq \|A^{-1}\| \|A - L\| < 1, \quad (5.8)$$

and thus $I - A^{-1}(A - L) \in \mathrm{GL}(X)$; see Proposition 5.7.4. By Remark 2.2.11 and Exercise 2.6, we have that $L \in \mathrm{GL}(X)$, so the claim holds and $\mathrm{GL}(X)$ is open.

To see the continuity of the inverse map, first note that for any invertible A and L , we have $L^{-1} = (A^{-1}L)^{-1}A^{-1}$, so if $\|L - A\| < \frac{1}{2\|A^{-1}\|}$, then

$$\begin{aligned} \|L^{-1} - A^{-1}\| &= \|(I - A^{-1}(A - L))^{-1}A^{-1} - A^{-1}\| \\ &= \left\| \left(\sum_{k=0}^{\infty} (A^{-1}(A - L))^k \right) A^{-1} - A^{-1} \right\| \\ &\leq \left(\sum_{k=1}^{\infty} \|A^{-1}(A - L)\|^k \right) \|A^{-1}\| \\ &= \frac{\|A^{-1}(A - L)\|}{1 - \|A^{-1}(A - L)\|} \|A^{-1}\| \\ &\leq \frac{\|A^{-1}\|^2 \|A - L\|}{1 - \|A^{-1}(A - L)\|}. \end{aligned} \quad (5.9)$$

Given $\varepsilon > 0$, set $\delta = \min(\frac{\varepsilon}{2\|A^{-1}\|^2}, \frac{1}{2\|A^{-1}\|})$, so that whenever $\|L - A\| < \delta$ we have

$$\frac{1}{1 - \|A^{-1}(A - L)\|} < 2,$$

and thus

$$\|L^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|^2 \|A - L\|}{1 - \|A^{-1}(A - L)\|} < 2\|A^{-1}\|^2 \|A - L\| < \varepsilon$$

whenever $\|L - A\| < \delta$. Hence, the map $A \mapsto A^{-1}$ is continuous. $\square$

The proof of the previous proposition actually gives some explicit bounds on the norm of inverses and the size of the open ball in $\mathrm{GL}(X)$ containing the inverse of some operator. This is useful for studying pseudospectra in Chapter 14.

Proposition 5.7.8. *Let $(X, \|\cdot\|_X)$ be a Banach space. Suppose $A \in \mathcal{B}(X)$ satisfies $\|A^{-1}\| < M$. For any $E \in \mathcal{B}(X)$ with $\|E\| < 1/\|A^{-1}\|$, the operator $A + E$ has a bounded inverse satisfying*

$$\|(A + E)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|E\|\|A^{-1}\|}.$$

Proof. The proof is Exercise 5.44. $\square$

5.8 Topologically Equivalent Metrics

Recall that for a metric space (X, d) the collection of all the open sets is called the *topology* induced by the metric d . Changing the metric generally changes the topology, but sometimes different metrics define the same topology. In that case, we say that the metrics are *topologically equivalent*. For example, in this section we show that any metric induced by a norm on $\mathbb{F}^n$ is topologically equivalent to the Euclidean metric.

Properties that depend only on open sets—topological properties—are the same for all topologically equivalent metrics. Some examples of topological properties that you have already encountered are continuity and compactness. In the following section we also discuss another important topological property called *connectedness*.

One reason that all this discussion of topology is helpful is that often it is easier to prove that a certain property holds for one metric than for another. If the metrics are topologically equivalent, and if the property in question is a topological one, then it is enough to prove the property holds for one metric, and that automatically implies it holds for the other.

5.8.1 Topological Equivalence

Definition 5.8.1. *Let X be a set with two metrics d_a and d_b . Each metric induces a collection of open sets, and this collection is called a topology. We denote the set of all d_a -open sets by $\mathcal{T}_a$ and call it the a -topology. Similarly, we denote the set of all d_b -open sets by $\mathcal{T}_b$ and call it the b -topology. We say that d_a and d_b are topologically equivalent if they define the same open sets on X , that is, if $\mathcal{T}_a = \mathcal{T}_b$.*

Example 5.8.2.

- (i) Let $\mathcal{T}$ be the topology induced on a set X by the discrete metric (see Example 5.1.3(iii)). For any point $x \in X$ and for any $r < 1$, we have $B(x, r) = \{x\}$. Since arbitrary unions of open sets are open, this means that any set is open with respect to this metric, and the topology on X induced by this metric is the entire power set 2^X of X .

- (ii) The 2-norm on $\mathbb{F}^n$ induces the usual Euclidean metric and the *Euclidean topology*. The open sets in this topology are what most mathematicians usually mean when they say “open” without any other statement about the metric (or the topology). In particular, open sets in $\mathbb{R}^1$ with the Euclidean topology are infinite unions and finite intersections of open intervals (a, b) .

Example 5.3.12 shows that the Euclidean topology on $\mathbb{F}^n$ is not topologically equivalent to the discrete topology.

- (iii) The space $C([0, 1]; \mathbb{R})$ with the sup norm defines a topology via the metric $d_\infty(f, g) = \|f - g\|_{L^\infty} = \sup_{x \in [0, 1]} |f(x) - g(x)|$.

We may also define the L^1 -norm and its corresponding metric as $d_1(f, g) = \|f - g\|_1 = \int_0^1 |f(t) - g(t)| dt$.

Every open set of the L^1 topology is also an open set of the sup-topology because for any $f \in C([0, 1]; \mathbb{R})$ and for any $\varepsilon > 0$, the ball $B_\infty(f, \varepsilon/2) \subset B_1(f, \varepsilon)$. To see this, observe that for any $g \in B_\infty(f, \varepsilon/2)$ we have $\|f - g\|_{L^1} = \int_0^1 |f - g| dt \leq \int_0^1 \varepsilon/2 dt = \varepsilon/2 < \varepsilon$.

In Unexample 5.8.6 we show that the L^1 -metric is *not* topologically equivalent to the L^∞ -metric.

5.8.2 Characterization of Topologically Equivalent Metrics

Theorem 5.8.3. *Let X be a set with two metrics, d_a and d_b . The metrics d_a and d_b are topologically equivalent if and only if for all $\mathbf{x} \in X$ and for all $\varepsilon > 0$ there exist $\delta_a, \delta_b > 0$ such that*

$$B_b(\mathbf{x}, \delta_b) \subset B_a(\mathbf{x}, \varepsilon) \quad \text{and} \quad B_a(\mathbf{x}, \delta_a) \subset B_b(\mathbf{x}, \varepsilon), \quad (5.10)$$

where B_a and B_b are the open balls defined by the metrics d_a and d_b , respectively.

Proof. If d_a and d_b are topologically equivalent, then every ball $B_a(\mathbf{x}, \varepsilon)$ is open with respect to d_b , and by the definition of open set (Definition 5.1.8) there must be a ball $B_b(\mathbf{x}, \delta)$ contained in $B_a(\mathbf{x}, \varepsilon)$. Similarly, every ball $B_b(\mathbf{x}, \varepsilon)$ is open with respect to d_a and there must be a ball $B_a(\mathbf{x}, \gamma)$ contained in $B_b(\mathbf{x}, \varepsilon)$.

Conversely, let $U \subset X$ be an open set in the metric space (X, d_a) . For each $\mathbf{x} \in U$ there exists $\varepsilon > 0$ such that $B_a(\mathbf{x}, \varepsilon) \subset U$, and by hypothesis there is a $\delta_b > 0$ such that $B_b(\mathbf{x}, \delta_b) \subset B_a(\mathbf{x}, \varepsilon) \subset U$. Hence, U is open in the metric space (X, d_b) . Interchanging the roles of a and b in this argument shows that every set that is open in (X, d_b) is also open in (X, d_a) . Hence, the topologies are equivalent. $\square$

5.8.3 Metrics Induced by Norms

If we restrict ourselves to normed linear spaces and only allow metrics induced by norms, then we can strengthen Theorem 5.8.3 as follows.

Theorem 5.8.4. *Let X be a vector space with two norms, $\|\cdot\|_a$ and $\|\cdot\|_b$. The metrics induced by these norms are topologically equivalent if and only if there exist constants $0 < m \leq M$ such that*

$$m\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq M\|\mathbf{x}\|_a \quad (5.11)$$

for all $\mathbf{x} \in X$.

Proof. ($\Rightarrow$) If (5.11) holds for some $\mathbf{x} \in X$, then it also holds for every scalar multiple of $\mathbf{x}$. Therefore, it suffices to prove that (5.11) holds for every $\mathbf{x} \in B(0, 1)$. Let d_a and d_b be the metrics on X induced by the norms $\|\cdot\|_a$ and $\|\cdot\|_b$, respectively. If d_a and d_b are topologically equivalent, then by Theorem 5.8.3 there exist $0 < \varepsilon' < 1$ and $0 < \varepsilon'' < 1$ such that

$$B_a(\mathbf{0}, \varepsilon'') \subset B_b(\mathbf{0}, \varepsilon') \subset B_a(\mathbf{0}, 1). \quad (5.12)$$

By Exercise 5.50 we have

$$\|\mathbf{x}\|_a \leq \frac{\|\mathbf{x}\|_b}{\varepsilon'} \leq \frac{\varepsilon'' \|\mathbf{x}\|_a}{\varepsilon'}.$$

Setting $m = \varepsilon'$ and $M = \varepsilon''$ gives (5.11).

($\Leftarrow$) If (5.11) holds for all $\mathbf{x} \in X$, then it also holds for all $(\mathbf{x} - \mathbf{y}) \in X$; that is, for all $\mathbf{x}, \mathbf{y} \in X$ we have

$$m\|\mathbf{x} - \mathbf{y}\|_a \leq \|\mathbf{x} - \mathbf{y}\|_b \leq M\|\mathbf{x} - \mathbf{y}\|_a.$$

Hence, $md_a(\mathbf{x}, \mathbf{y}) \leq d_b(\mathbf{x}, \mathbf{y})$ and $d_b(\mathbf{x}, \mathbf{y}) \leq Md_a(\mathbf{x}, \mathbf{y})$, which implies that for every $\varepsilon > 0$ we have

$$B_a(\mathbf{x}, m\varepsilon) \subset B_b(\mathbf{x}, \varepsilon) \quad \text{and} \quad B_b(\mathbf{x}, \varepsilon/M) \subset B_a(\mathbf{x}, \varepsilon).$$

Therefore, by Theorem 5.8.3 (with $\delta_a = m\varepsilon$ and $\delta_b = \varepsilon/M$) the two norms are topologically equivalent. $\square$

Example 5.8.5. Recall the following inequalities on $\mathbb{F}^n$ from Exercise 3.17:

$$(i) \quad \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2.$$

$$(ii) \quad \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty.$$

This shows that these norms are topologically equivalent on $\mathbb{F}^n$.

Unexample 5.8.6. The norms L^∞ and L^1 are not topologically equivalent on $C([0, 1]; \mathbb{R})$. To see this, let $f_n(x) = x^n$ for each $n \in \mathbb{N}$. It is straightforward to check that $\|f_n\|_{L^\infty} = 1$ and $\|f_n\|_{L^1} = \frac{1}{n+1}$, which shows that (5.11) cannot hold.

Theorem 5.8.7. *All norms on a finite-dimensional vector space are topologically equivalent.*

Proof. Let $\|\cdot\|$ be a given norm on a finite-dimensional vector space X . By Corollary 2.3.12 there is a vector-space isomorphism $f : X \rightarrow \mathbb{F}^n$ for some $n \in \mathbb{N}$. The map f induces a norm $\|\cdot\|_f$ on $\mathbb{F}^n$ by $\|\mathbf{x}\|_f = \|f^{-1}(\mathbf{x})\|$ (verification that this is a norm is Exercise 5.48). Therefore, it suffices to assume that $X = \mathbb{F}^n$.

Moreover, it suffices to show that $\|\cdot\|$ on $\mathbb{F}^n$ is topologically equivalent to the 1-norm $\|\cdot\|_1$. From Example 5.5.8, we know that the unit sphere in the 1-norm is compact and that the function $f(\mathbf{x}) = \|\mathbf{x}\|$ is continuous in the 1-norm topology, so its image contains both its maximum M and its minimum m . Thus, for every nonzero $\mathbf{x} \in \mathbb{F}^n$ we have $m \leq \frac{\|\mathbf{x}\|}{\|\mathbf{x}\|_1} \leq M$, hence $m\|\mathbf{x}\|_1 \leq \|\mathbf{x}\| \leq M\|\mathbf{x}\|_1$. $\square$

Remark 5.8.8. The previous theorem does not hold for infinite-dimensional spaces, as shown in Unexample 5.8.6

Remark 5.8.9. Recall the Cartesian product of several metric spaces $X_1 \times \cdots \times X_n$ described in Example 5.1.3(iv). Following the approach of Theorem 5.8.7, one can show that for any $p, q \in [1, \infty)$ the p -metric is topologically equivalent to the q -metric on $X_1 \times \cdots \times X_n$.

5.9 Topological Properties

A property of metric spaces is *topological* if it depends only on the topology of the space and not on the metric. In other words, a property is topological if whenever it holds for one metric on X , it also holds for any topologically equivalent metric on X . Any time we are interested in studying a topological property, we may switch from one metric to any topologically equivalent metric. Depending on the situation, some metrics may be much easier to work with than others, so this can simplify many arguments and computations. In particular, Theorem 5.8.7 tells us that any time we are interested in studying a topological property of a finite-dimensional normed linear space, we can use whichever norm is the easiest to work with for that problem.

5.9.1 Homeomorphisms

To begin our study of topological properties, we first turn to continuous functions. Continuous functions are important in this setting because the preimage of an open set is open (see Theorem 5.2.3). Therefore, a function that is continuous and has a continuous inverse preserves all topological properties. We use this fact to make the idea of a topological property more precise.

Definition 5.9.1. *Let (X, d) and (Y, ρ) be metric spaces. A homeomorphism $f : (X, d) \rightarrow (Y, \rho)$ is a bijective, continuous map whose inverse f^{-1} is also continuous.*

Example 5.9.2. The map $f : (0, 1) \rightarrow (1, \infty) \subset \mathbb{R}$ given by $f(t) = 1/t$ is a homeomorphism: it is clearly bijective and continuous, and its inverse is $f^{-1}(s) = 1/s$, which is also continuous.

Unexample 5.9.3. A bijective, continuous function need not be a homeomorphism. For example, the map $f : [0, 1) \rightarrow S^1 = \{z \in \mathbb{C} \mid 1 = |z|\}$ given by $f(t) = e^{2\pi it}$ is continuous and bijective, but its inverse is not continuous at $z = 1$. This can be seen by the fact that S^1 is compact, but $[0, 1)$ is not. Since continuous functions must map compact sets to compact sets (see Proposition 5.5.6), the map f^{-1} cannot be continuous.

Definition 5.9.4. A topological property is one that is preserved under homeomorphism.

Example 5.9.5. Open sets are preserved under homeomorphism, as are closed sets. Compactness is defined only in terms of open sets, so it is also preserved under homeomorphism. Theorem 5.3.19 guarantees that convergence of sequences is preserved by continuous functions, and therefore convergence is a topological property.

Proposition 5.9.6. Two metrics d and ρ on X are topologically equivalent if and only if the identity map $i : (X, d) \rightarrow (X, \rho)$ is a homeomorphism.

Proof. The proof is Exercise 5.51. $\square$

Corollary 5.9.7. Let (X, d) be a metric space. If ρ is another metric on X that is topologically equivalent to d , then a sequence $(\mathbf{x}_n)_{n=0}^\infty$ in X converges to $\mathbf{x}$ in (X, d) if and only if it converges to $\mathbf{x}$ in (X, ρ) .

Proof. By Proposition 5.9.6 the identity map $i : (X, d) \rightarrow (X, \rho)$ is a homeomorphism, and the result now follows because convergence and limits are preserved by homeomorphisms (see Theorem 5.3.19). $\square$

5.9.2 Topological versus Uniform Properties

Completeness is not a topological property (see Example 5.4.10), and a sequence that is Cauchy in one metric space (X, d) is not necessarily Cauchy for a topologically equivalent metric (X, ρ) . However, these are *uniform* properties, meaning that they are preserved by uniformly continuous functions (see Theorem 5.4.24). Remarkably, the next proposition shows that if two metrics d_a, d_b on a vector space X are induced by *topologically equivalent norms* $\|\cdot\|_a, \|\cdot\|_b$, then the identity

map $i : (X, d_a) \rightarrow (X, d_b)$ is *uniformly* continuous (and its inverse is also uniformly continuous), so Cauchy sequences and completeness are preserved by topologically equivalent norms.

Proposition 5.9.8. *If $(X, \|\cdot\|_a)$ is a normed linear space, and if $\|\cdot\|_b$ is another norm on X that is topologically equivalent to $\|\cdot\|_a$, then the identity map $i : (X, \|\cdot\|_a) \rightarrow (X, \|\cdot\|_b)$ is uniformly continuous with a uniformly continuous inverse.*

Proof. It is immediate from the definition of topologically equivalent norms that the identity map is a bounded linear transformation, and hence by Proposition 5.4.22 it is uniformly continuous. The same applies to its inverse. $\square$

Remark 5.9.9. It is important to note that the previous result only holds for normed linear spaces and metrics induced by norms. It does not hold for more general metrics.

5.9.3 Connectedness

Connectedness is an important topological property. You might think the idea is obvious—a connected space shouldn't break into separate parts and you should be able to get from any point to any other by traveling in the space. But to actually prove anything we need to make these ideas mathematically precise. Being precise about definitions also reveals that these are actually two distinct ideas. Connectedness (not being able to break the space apart) is not quite the same as path connectedness (being able to get from any point to any other along a path in the space).

We begin this section with the careful definition of connectedness. We then discuss some of the consequences of connectedness and conclude with a discussion of path connectedness.

Connectedness

Definition 5.9.10. *A metric space X is disconnected if there are disjoint nonempty open subsets U and V such that $X = U \cup V$. In this case, we say the subsets U and V disconnect X . If X is not disconnected, then it is connected.*

Remark 5.9.11. If X is disconnected, then we can choose disjoint open sets U and V with $X = U \cup V$. This implies that $U = V^c$ and $V = U^c$ are also closed. Hence, a space X is connected if and only if the only sets that are both open and closed are X and $\emptyset$.

Example 5.9.12.

- (i) As we see later in this section, the line $\mathbb{R}$ is connected. But the set $\mathbb{R} \setminus \{0\}$ is disconnected because it is the union of the two disjoint open sets $(-\infty, 0)$ and $(0, \infty)$.

- (ii) If X is any set with at least two points and d is the discrete metric on X , then (X, d) is disconnected. This is because every set in the discrete topology is both open and closed; see Example 5.8.2(i).

Theorem 5.9.13. *Let $f : (X, d) \rightarrow (Y, \rho)$ be continuous and surjective. If X is connected, then so is Y .*

Proof. Suppose Y is not connected. There must be nonempty disjoint open sets U and V satisfying $Y = U \cup V$. Since f is continuous and surjective, the sets $f^{-1}(U)$ and $f^{-1}(V)$ are also nonempty disjoint open sets satisfying $X = f^{-1}(U) \cup f^{-1}(V)$. This is a contradiction. $\square$

Consequences of connectedness

Our first important consequence of connectedness is a corollary of Theorem 5.9.13.

Corollary 5.9.14 (Intermediate Value Theorem). *Assume (X, d) is a connected metric space, and let $f : (X, d) \rightarrow \mathbb{R}$ be continuous (with the usual topology on $\mathbb{R}$). If $f(\mathbf{x}) < f(\mathbf{y})$ and $c \in (f(\mathbf{x}), f(\mathbf{y}))$, then there exists $\mathbf{z} \in X$ such that $f(\mathbf{z}) = c$.*

Proof. If no such $\mathbf{z}$ exists, then $f^{-1}((-\infty, c))$ and $f^{-1}((c, \infty))$ are nonempty and disconnect X , which is a contradiction. $\square$

Example 5.9.15. Similar to the proof of the intermediate value theorem, we can use connectedness to show that an odd-degree polynomial $p \in \mathbb{R}[x]$ has a real root. If not, then $p^{-1}((-\infty, 0))$ and $p^{-1}((0, \infty))$ are both nonempty and they disconnect $\mathbb{R}$, which contradicts the fact that $\mathbb{R}$ is a connected set. But if p has even degree, then $p^{-1}((-\infty, 0))$ or $p^{-1}((0, \infty))$ may be empty, so this argument does not hold for even-degree polynomials.

The intermediate value theorem leads to our first example of a *fixed-point theorem*. Fixed-point theorems are used throughout mathematics and are also widely used in economics. We encounter them again in Chapter 7.

Corollary 5.9.16 (One-Dimensional Brouwer Fixed-Point Theorem). *If $f : [a, b] \rightarrow [a, b]$ is continuous, then there exists $x \in [a, b]$ such that $f(x) = x$.*

Proof. We assume that $f(a) > a$ and $f(b) < b$, otherwise $f(a) = a$ or $f(b) = b$, and the conclusion follows. Hence, the function $g(x) = x - f(x)$ satisfies $g(a) < 0$ and $g(b) > 0$, and is continuous on $[a, b]$. Thus, by the intermediate value theorem, there exists $c \in (a, b)$ such that $g(c) = 0$, or in other words, $f(c) = c$. $\square$

An illustration of the Brouwer fixed-point theorem is given in Figure 5.3.

Corollary 5.9.17 (One-Dimensional Borsuk–Ulam Theorem). *A continuous real-valued map f on the unit circle $S^1 = \{(x, y) \in \mathbb{R}^2 : \|(x, y)\|_2 = 1\}$ has antipodal points that are equal; that is, there exists $\mathbf{z} \in S^1$ such that $f(\mathbf{z}) = f(-\mathbf{z})$.*

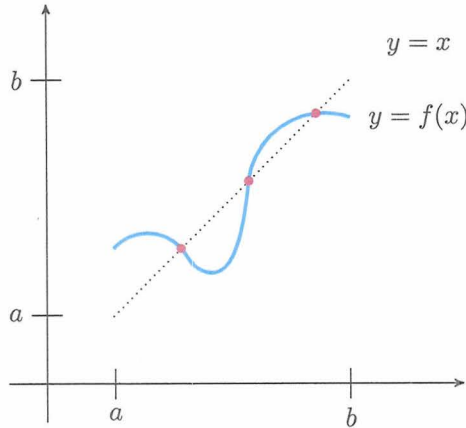


Figure 5.3. Plot of a continuous function $f : [a, b] \rightarrow [a, b]$. The Brouwer fixed-point theorem (Corollary 5.9.16) guarantees that any such function must have at least one fixed point, where $f(x) = x$. In this particular case f has three fixed points (red).

Proof. The proof is Exercise 5.54. $\square$

Remark 5.9.18. Connectedness is a very powerful, yet subtle, property. It implies, for example, by way of Corollary 5.9.17, that at any point in time there are antipodal points on the earth (or on any circle on the surface of the earth) that are exactly the same temperature!

Path connectedness

Connectedness is defined in terms of not being able to separate a space, but as mentioned in the introduction to this section, it may seem intuitive that you should be able to get from any point in a connected space to any other point by traveling within the space. This second property is actually stronger than the definition of connectedness. We call this stronger property *path connectedness*.

Definition 5.9.19. A subset E of a metric space is path connected if for any $\mathbf{x}, \mathbf{y} \in E$ there is a continuous map $\gamma : [0, 1] \rightarrow E$ such that $\gamma(0) = \mathbf{x}$ and $\gamma(1) = \mathbf{y}$. Such a γ is called a path from $\mathbf{x}$ to $\mathbf{y}$. See Figure 5.4.

To understand path-connected spaces, we first need to know that the interval $[0, 1]$ is connected.

Proposition 5.9.20. The interval $[0, 1] \subset \mathbb{R}$ is connected (in the usual topology).

Proof. Suppose that the interval has a separating pair $U \cup V = [0, 1]$. Take $u \in U$ and $v \in V$. Without loss of generality, assume $u < v$. The interval $[u, v]$ is a subset of $[0, 1]$, so it is contained in $U \cup V$. The set $A = [u, v] \cap U = [u, v] \cap V^c$ is compact, because it is a closed subset of the compact interval $[u, v]$. Therefore A contains its largest element a . Also $a < v$ because $v \notin U$.

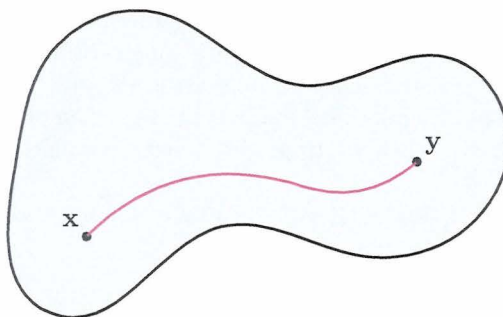


Figure 5.4. In a path-connected space, there is a path within the space between any two points; see Definition 5.9.19.

Similarly, the least element b of $[a, v] \cap V$ is contained in $[a, v] \cap V$. Since $b \notin A$ we must have $a < b$, which shows that the interval $(a, b) \subset [u, v]$ is not empty. But $(a, b) \cap U = (a, b) \cap V = \emptyset$; hence, not every element of $[0, 1]$ is contained in $U \cup V$, a contradiction. $\square$

Theorem 5.9.21. A path-connected space is connected.³⁷

Proof. Assume that X is path connected but not connected. Hence, there exists a pair of nonempty disjoint open sets U and V such that $X = U \cup V$. Choose $x \in U$ and $y \in V$ and a continuous map $\gamma : [0, 1] \rightarrow X$ such that $\gamma(0) = x$ and $\gamma(1) = y$. Thus, the sets $\gamma^{-1}(U)$ and $\gamma^{-1}(V)$ are nonempty, disjoint, and disconnect the connected set $[0, 1]$, which is a contradiction. $\square$

Example 5.9.22. Any interval in $\mathbb{R}$ is path connected, and thus connected.

5.10 Banach-Valued Integration

In this section, we use the continuous linear extension theorem to construct the integral of a single-variable Banach-valued function on a compact interval. To do this we first define the integral on some of the simplest functions imaginable—step functions. For step functions the definition of the integral is very easy and obvious. Since continuous functions are in the closure (under the sup norm $\|\cdot\|_{L^\infty}$) of the space of step functions, this means that once we show that the integral, as a linear operator, is bounded, we can use the continuous linear extension theorem to extend the integral to all continuous functions. These integrals are used extensively in Chapter 6. In particular, we use these integrals to define Taylor series in a Banach space (see Section 6.6).

³⁷The converse to this theorem is false. There are spaces that are connected but not path connected. For an example of such a space, see [Mun75, Ex. 2, Sect. 25].

An immediate consequence of the fact that the integral is a continuous linear operator is that uniform limits commute with integration. With all the tools we now have at our disposal, it is easy to define the integral and prove many of its properties—much easier than with the traditional definition of the Riemann integral. We use these tools again when we treat the Lebesgue construction of the integral in Chapter 8.

Throughout this section, let $(X, \|\cdot\|_X)$ be a Banach space.

Nota Bene 5.10.1. The approach we use to define the integral is unusual. Aside from Dieudonné's books [Bou76, Die60], we know of no textbook that treats the integral this way. But it has many advantages over the more common approaches, including being much simpler. We believe it is also a better preparation for the ideas of Lebesgue integration. The construction of the integral given here is called the *regulated integral*, and although it is similar to the Riemann integral in many ways, it is not identical to it. Nevertheless, it is straightforward to show that whenever the two constructions are both defined, they must agree.

5.10.1 Integrals of Step Functions

Definition 5.10.2. A map $f : [a, b] \rightarrow X$ is a step function if there is a (finite) subdivision $a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$ such that we may write f in the form

$$f = \left(\sum_{i=1}^{N-1} \mathbf{x}_i \mathbb{1}_{[t_{i-1}, t_i)} \right) + \mathbf{x}_N \mathbb{1}_{[t_{N-1}, t_N]}, \quad (5.13)$$

where each $\mathbf{x}_i \in X$ and $\mathbb{1}_E$ is the indicator function of the set E :

$$\mathbb{1}_E(t) = \begin{cases} 1 & \text{if } t \in E, \\ 0 & \text{if } t \notin E. \end{cases} \quad (5.14)$$

Let $S([a, b]; X)$ denote the set of all step functions mapping $[a, b]$ into X .

For an illustration of a step function see Figure 5.5.

Proposition 5.10.3. The set $S([a, b]; X)$ of step functions is a subspace of the normed linear space of bounded functions $(L^\infty([a, b]; X), \|\cdot\|_{L^\infty})$.

Proof. To see that $S([a, b]; X)$ is a subset of $L^\infty([a, b]; X)$, note that a step function f of the form (5.13) has finite sup norm since $\|f\|_{L^\infty} = \sup_k \|\mathbf{x}_k\|_X$.

It suffices to show that $S([a, b]; X)$ is closed under linear combinations; that is, given $\alpha, \beta \in \mathbb{F}$ and $f, g \in S([a, b]; X)$, we show that $\alpha f + \beta g \in S([a, b]; X)$. Let

$$f(t) = \left(\sum_{i=1}^{N-1} \mathbf{x}_i \mathbb{1}_{[t_{i-1}, t_i)} \right) + \mathbf{x}_N \mathbb{1}_{[t_{N-1}, t_N]}$$

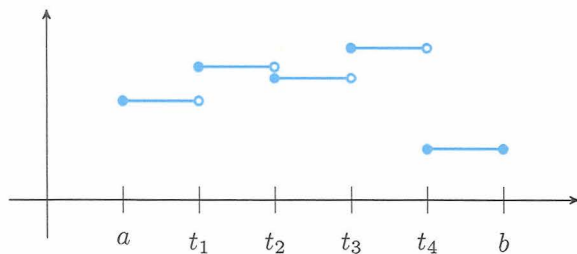


Figure 5.5. Graph of a step function $s : [a, b] \rightarrow \mathbb{R}$, as described in Definition 5.10.2.

and

$$g(t) = \left(\sum_{i=1}^{M-1} \mathbf{y}_i \mathbb{1}_{[s_{i-1}, s_i)} \right) + \mathbf{y}_M \mathbb{1}_{[s_{M-1}, s_M]}.$$

Write the union of the indices $\{t_0, \dots, t_N, s_0, \dots, s_M\}$ as an ordered list $u_0 < u_1 < \dots < u_\ell$ (first eliminating any duplicates). The step functions f and g can be rewritten as

$$f(t) = \left(\sum_{i=1}^{\ell-1} \mathbf{x}'_i \mathbb{1}_{[u_{i-1}, u_i)} \right) + \mathbf{x}'_\ell \mathbb{1}_{[u_{\ell-1}, u_\ell]} \text{ and } g(t) = \left(\sum_{i=1}^{\ell-1} \mathbf{y}'_i \mathbb{1}_{[u_{i-1}, u_i)} \right) + \mathbf{y}'_\ell \mathbb{1}_{[u_{\ell-1}, u_\ell]},$$

where each $\mathbf{x}'_i$ and $\mathbf{y}'_i$ lies in X . Thus, the sum takes the form

$$\alpha f(t) + \beta g(t) = \left(\sum_{i=1}^{\ell-1} (\alpha \mathbf{x}'_i + \beta \mathbf{y}'_i) \mathbb{1}_{[u_{i-1}, u_i)} \right) + (\alpha \mathbf{x}'_\ell + \beta \mathbf{y}'_\ell) \mathbb{1}_{[u_{\ell-1}, u_\ell]}.$$

This is a step function on $[a, b]$. $\square$

Definition 5.10.4. The integral of a step function $f \in S([a, b]; X)$ of the form (5.13) is defined to be

$$I(f) = \sum_{i=1}^N \mathbf{x}_i (t_i - t_{i-1}). \quad (5.15)$$

This is a map from $S([a, b]; X)$ to X . We often write $I(f)$ as $\int_a^b f(t) dt$.

Proposition 5.10.5. The integral map $I : S([a, b]; X) \rightarrow X$ is a bounded linear transformation with induced norm $\|I\| = (b - a)$.

Proof. Linearity follows from the combining of two subdivisions, as described in the proof of Proposition 5.10.3. For any step function f of the form (5.13) on $[a, b]$, we have

$$\|I(f)\| = \left\| \sum_{i=1}^N \mathbf{x}_i (t_i - t_{i-1}) \right\| \leq (b - a) \sup_k \|\mathbf{x}_k\| = (b - a) \|f\|_{L^\infty},$$

which gives $\|I\| \leq (b - a)$. But if $f(t) = \mathbb{1}_{[a,b]}(t)$, then $\|I(f)\| = (b - a) = (b - a)\|f\|_{L^\infty}$, and hence $\|I\| = (b - a)$. $\square$

5.10.2 Single-Variable Banach-Valued Integration

Theorem 5.10.6 (Single-Variable Banach-Valued Integration). *Let $L^\infty([a, b]; X)$ be given the L^∞ -norm. The linear map $I : S([a, b]; X) \rightarrow X$ can be extended uniquely to a bounded linear (hence uniformly continuous) transformation*

$$\bar{I} : \overline{S([a, b]; X)} \rightarrow X,$$

with $\|\bar{I}\| = \|I\| = (b - a)$.

Proof. This follows immediately from the continuous linear extension theorem (Theorem 5.7.6) by setting $S = S([a, b]; X)$ and $Z = \overline{S([a, b]; X)}$, since bounded linear transformations are always uniformly continuous, by Proposition 5.4.22. $\square$

Definition 5.10.7. *For any Banach space X and any function $f \in \overline{S([a, b]; X)}$, we write*

$$\int_a^b f(t) dt$$

to denote the unique linear extension $\bar{I}(f)$ of Theorem 5.10.6. In other words,

$$\int_a^b f(t) dt = \bar{I}(f) = \lim_{n \rightarrow \infty} I(s_n) = \lim_{n \rightarrow \infty} \int_a^b s_n(t) dt,$$

where $(s_n)_{n \in \mathbb{N}}$ is a sequence of step functions that converges uniformly to f . We also define

$$\int_b^a f(t) dt = - \int_a^b f(t) dt.$$

Theorem 5.10.8. *Continuous functions lie in the closure, with respect to the uniform norm, of the space of step functions*

$$C([a, b]; X) \subset \overline{S([a, b]; X)} \subset L^\infty([a, b]; X).$$

Proof. Any $f \in C([a, b]; X)$ is uniformly continuous by Theorem 5.5.9. So, given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|f(s) - f(t)\| < \varepsilon$ whenever $|s - t| < \delta$. Choose n sufficiently large so that $(b - a)/n < \delta$, and define a step function $f_s \in S([a, b]; X)$ as

$$f_s(t) = f(t_0)\mathbb{1}_{[t_0, t_1)}(t) + f(t_1)\mathbb{1}_{[t_1, t_2)}(t) + \cdots + f(t_{n-1})\mathbb{1}_{[t_{n-1}, t_n]}(t), \quad (5.16)$$

where the grid points $t_i = i(b - a)/n + a$ are equally spaced and less than δ distance apart. Since $t \in [t_i, t_{i+1}]$ implies that $|f(t) - f(t_i)| < \varepsilon$, we have that $\|f - f_s\|_{L^\infty} < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we have that f is a limit point of $S([a, b]; X)$. Since $f \in C([a, b]; X)$ was arbitrary, this shows that $C([a, b]; X) \subset \overline{S([a, b]; X)}$. $\square$

Remark 5.10.9. In the proof above we approximated a continuous function f with a step function that matched the values of f on the leftmost point of each subinterval. However, we could just as well have chosen any point on the subinterval to approximate the function f .

Remark 5.10.10. The tools we have built up throughout the book so far made our construction of the integral much simpler than the traditional (Riemann or Darboux) construction of the integral. We should point out that the Riemann construction applies to more functions than just $\overline{S([a, b]; X)}$, but in Chapter 8 we describe another construction (the Lebesgue or Daniell construction) that applies to even more functions than the Riemann construction. We use the continuous linear extension theorem as the main tool in that construction as well.

Remark 5.10.11. We won't prove it here, but it is straightforward to check that the usual Riemann construction of the integral also gives a bounded linear transformation from $\overline{S([a, b]; \mathbb{R})}$ to $\mathbb{R}$, and it agrees with our regulated integral on all real-valued step functions, so by the uniqueness of linear extensions, the Riemann and regulated constructions of the integral must be the same on $\overline{S([a, b]; \mathbb{R})}$. In particular, they must agree on all continuous functions. Among other things, this means that for functions in $\overline{S([a, b]; \mathbb{R})}$ (real valued) we can use any of the usual techniques of Riemann integration to evaluate the integral, including, for example, the fundamental theorem of calculus, change of variables (substitution), and integration by parts.

Of course the integral as we have defined it here is not limited to real-valued functions—it works just as well on functions that take values in any Banach space, and in that more general setting we cannot always use the usual techniques of Riemann integration. But we prove a Banach-valued fundamental theorem of calculus in Section 6.5.

5.10.3 Properties of the Integral

The fact that the extension $\overline{I}$ is linear means that if $f, g \in \overline{S([a, b]; X)}$ and $\alpha, \beta \in \mathbb{F}$, then

$$\int_a^b \alpha f(t) + \beta g(t) dt = \alpha \int_a^b f(t) dt + \beta \int_a^b g(t) dt.$$

The next proposition gives several more fundamental properties of integration.

Proposition 5.10.12. *If $f \in \overline{S([a, b]; X)} \subset L^\infty([a, b]; X)$ and $\alpha, \beta, \gamma \in [a, b]$, with $\alpha < \gamma < \beta$, then the following hold:*

- (i) $\|\int_a^b f(t) dt\| \leq (b - a) \sup_{t \in [a, b]} \|f(t)\|.$
- (ii) $\|\int_a^b f(t) dt\| \leq \int_a^b \|f(t)\| dt$ (integral triangle inequality).
- (iii) *Restricting f to a subinterval $[\alpha, \beta] \subset [a, b]$ defines a function that we also denote by $f \in \overline{S([\alpha, \beta]; X)}$. We have $\int_a^b f(t) \mathbb{1}_{[\alpha, \beta]}(t) dt = \int_\alpha^\beta f(t) dt.$*
- (iv) $\int_\alpha^\beta f(t) dt = \int_\alpha^\gamma f(t) dt + \int_\gamma^\beta f(t) dt.$
- (v) *The function $F(t) = \int_a^t f(s) ds$ is continuous on $[a, b].$*

Proof.

- (i) This is just a restatement of $\|\bar{I}\| = (b - a)$ on the space $\overline{S([a, b]; X)}$ in Proposition 5.10.5.
- (ii) The function $t \mapsto \|f(t)\|$ is an element of $\overline{S([a, b]; \mathbb{R})}$, so the integral $\int_a^b \|f(t)\| dt$ makes sense. The rest of the proof is Exercise 5.55.
- (iii) Let $(s_k)_{k=0}^\infty$ be a sequence of step functions in $S([a, b]; X)$ that converges to $f \in \overline{S([a, b]; X)}$ in the sup norm. Let I_{ab} be the integral map on $S([a, b]; X)$ as given in (5.15), and let $I_{\alpha\beta}$ be the corresponding integral map on $S([\alpha, \beta]; X)$. Since the product $s_k \mathbb{1}_{[\alpha, \beta]}$ vanishes outside of $[\alpha, \beta]$, its map for $I_{\alpha\beta}$, with restricted domain, has the same value as that of I_{ab} . In other words, for all $k \in \mathbb{N}$ we have

$$\int_a^b s_k(t) \mathbb{1}_{[\alpha, \beta]} dt = \int_\alpha^\beta s_k(t) dt.$$

Since $\bar{I}$ is continuous, $s_k \rightarrow f$ implies that $\bar{I}(s_k) \rightarrow \bar{I}(f)$, by Theorem 5.2.9.

- (iv) The proof is Exercise 5.56.

- (v) The proof is Exercise 5.57. $\square$

Remark 5.10.13. We can combine the results of (i) and (ii) in Proposition 5.10.12 into the following:

$$\left\| \int_a^b f(t) dt \right\| \leq \int_a^b \|f(t)\| dt \leq (b - a) \sup_{t \in [a, b]} \|f(t)\|.$$

In other words, the result in (ii) is a sharper inequality than that of (i).

Proposition 5.10.14. If $\mathbf{f} \in \overline{S([a, b]; \mathbb{R}^n)} \subset L^\infty([a, b], \mathbb{R}^n)$ is written in coordinates as $\mathbf{f}(t) = (f_1(t), \dots, f_n(t))$, then for each i we have $f_i \in \overline{S([a, b]; \mathbb{R})}$ and

$$\int_a^b \mathbf{f}(t) dt = \left(\int_a^b f_1(t) dt, \dots, \int_a^b f_n(t) dt \right).$$

Proof. Since $\bar{I} : \overline{S([a, b]; X)} \rightarrow X$ is continuous, if $s_n \rightarrow \mathbf{f}$, then $\bar{I}(s_n) \rightarrow \bar{I}(\mathbf{f})$ by Theorem 5.2.9. Thus, it suffices to prove the proposition for step functions. But for a step function $\mathbf{s}(t) = (s_1(t), \dots, s_n(t))$ the proof is straightforward:

$$\begin{aligned} \int_a^b \mathbf{s}(t) dt &= \sum_{i=0}^{n-1} \mathbf{s}(t_i)(t_i - t_{i-1}) \\ &= \sum_{i=0}^{n-1} (s_1(t_i), \dots, s_n(t_i))(t_i - t_{i-1}) \\ &= \left(\sum_{i=0}^{n-1} s_1(t_i)(t_i - t_{i-1}), \dots, \sum_{i=0}^{n-1} s_n(t_i)(t_i - t_{i-1}) \right) \\ &= \left(\int_a^b s_1(t) dt, \dots, \int_a^b s_n(t) dt \right). \quad \square \end{aligned}$$

Application 5.10.15. Integrals of single-variable continuous functions show up in many applications. One common application is in the physics of particle motion. If a particle moving in $\mathbb{R}^n$ has acceleration $\mathbf{a}(t) = (a_1(t), \dots, a_n(t))$, then its velocity at time t is $\mathbf{v}(t) = \int_{t_0}^t \mathbf{a}(\tau) d\tau + \mathbf{v}(t_0)$, and its position is $\mathbf{p}(t) = \int_{t_0}^t \mathbf{v}(\tau) d\tau + \mathbf{p}(t_0)$.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

5.1. If d_1 and d_2 are two given metrics on X , decide which of the following are metrics on X , and prove your answers are correct.

- (i) $d_1 + d_2$.
- (ii) $d_1 - d_2$.
- (iii) $\min(d_1, d_2)$.
- (iv) $\max(d_1, d_2)$.

5.2. Let d_2 denote the usual Euclidean metric on $\mathbb{R}^2$. Let $\mathbf{0} = (0, 0)$ be the origin. The *French railway metric* d_{SNCF} in $\mathbb{R}^2$ is given by³⁸

$$d_{\text{SNCF}}(\mathbf{x}, \mathbf{y}) = \begin{cases} d_2(\mathbf{x}, \mathbf{y}), & \mathbf{x} = \alpha \mathbf{y} \text{ for some } \alpha \in \mathbb{R}, \\ d_2(\mathbf{x}, \mathbf{0}) + d_2(\mathbf{0}, \mathbf{y}) & \text{otherwise.} \end{cases}$$

Explain why *French railway* might be a good name for this metric. (Hint: Think of Paris as the origin.) Prove that d_{SNCF} really is a metric on $\mathbb{R}^2$. Describe the open balls $B(\mathbf{x}, \varepsilon)$ in this metric.

³⁸The official name of the French national railway is *Société Nationale des Chemins de Fer*, or SNCF for short.

- 5.3. Give an example of a set X and a function $d : X \times X \rightarrow \mathbb{R}$ that is
- (i) symmetric and satisfies the triangle inequality, but is not positive definite;
 - (ii) positive definite and symmetric, but does not satisfy the triangle inequality.
- 5.4. Let (X, d) be a metric space. Show that the function (5.5) of Example 5.1.4 is also a metric. Hint: The function $f(x) = \frac{x}{1+x}$ is monotonically increasing on $[0, \infty)$.
- 5.5. For all $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in a metric space (X, d) , prove that $d(\mathbf{x}, \mathbf{z}) \geq |d(\mathbf{x}, \mathbf{y}) - d(\mathbf{y}, \mathbf{z})|$.
- 5.6. Prove Theorem 5.1.16(iii).
- 5.7.* Consider the collection $\mathcal{C}$ in Example 5.1.13. Prove that every point in the unit ball $B(0, 1)$ (with respect to the usual Euclidean metric) is contained in one of the elements of $\mathcal{C}$. Hint: $\mathbf{x} \in B(0, 1)$ if and only if $\mathbf{x} = \sum c_i \mathbf{e}_i$, where $\sum c_i^2 < 1$. Show that $\|\mathbf{x} - \mathbf{e}_i\| < \sqrt{2}$ for some $\mathbf{e}_i$.
- 5.8.* Let $((X_k, d_k))_{k=0}^\infty$ be an infinite sequence of metric spaces. Show that the function

$$d(\mathbf{x}, \mathbf{y}) = \sum_{k=0}^{\infty} \frac{1}{2^k} \cdot \frac{d_k(\mathbf{x}, \mathbf{y})}{1 + d_k(\mathbf{x}, \mathbf{y})} \quad (5.17)$$

is a metric on the Cartesian product $\prod_{k=0}^\infty X_k$. Note: Remember to show that the sum always converges to a finite value.

- 5.9. Let (X, d) be a metric space, and let $B \subset X$ be nonempty. For $\mathbf{x} \in X$, define

$$\rho(B, \mathbf{x}) = \inf\{d(\mathbf{x}, \mathbf{b}) \mid \mathbf{b} \in B\}.$$

Show that, for a fixed B , the function $\rho(B, \cdot) : X \rightarrow \mathbb{R}$ is a continuous function.

- 5.10. Prove that multivariable polynomials over $\mathbb{F}$ are continuous. Hint: Use Proposition 5.2.11.
- 5.11. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be

$$f(x, y) = \begin{cases} 0 & \text{if } x = y = 0, \\ \frac{x^2 - y^2}{\sqrt{x^2 + y^2}} & \text{otherwise.} \end{cases}$$

Prove that f is continuous at $(0, 0)$ (using the Euclidean metric $d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|_2$ in $\mathbb{R}^2$, and the usual metric $d(a, b) = |a - b|$ in $\mathbb{R}$).

- 5.12. Let

$$f(x, y) = \begin{cases} 0 & \text{if } x = y = 0, \\ \frac{2x^2y}{x^4 + y^2} & \text{otherwise.} \end{cases}$$

Define $\phi(t) = (t, at)$ and $\psi(t) = (t, t^2)$. Show that

- (i) $\lim_{t \rightarrow 0} f(\phi(t)) = 0$;
- (ii) $\lim_{t \rightarrow 0} f(\psi(t)) = 1$.

What does this say about the continuity of f ? Explain your results.

- 5.13. Prove the claim made in Example 5.2.14 that $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ is the set of all limit points of the ball $B(\mathbf{0}, 1) \subset \mathbb{R}^2$.
- 5.14. Prove that for each integer $n > 0$, the set $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$ in the usual (Euclidean) metric.

- 5.15. Prove that a subset $E \subset X$ is dense in X if and only if $\overline{E} = X$.
- 5.16. Prove that $(\overline{E})^c = (E^c)^\circ$.
- 5.17. Prove that $\mathbf{x}_0 \in \overline{E}$ if and only if $\inf_{\mathbf{x} \in E} d(\mathbf{x}_0, \mathbf{x}) = 0$.
- 5.18. Give an example of a metric space (X, d) and two proper (nonempty and not equal to X) subsets $S, T \subset X$ such that S is both open and closed, and T is neither open nor closed.
- 5.19.  Prove that the kernel of a bounded linear transformation is closed. Hint: Consider using Example 5.2.2(ii).
- 5.20. For each of the following functions $f : \mathbb{R}^2 \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}$, determine whether the limit of $f(x, y)$ exists at $(0, 0)$. Hint: If the limit exists, then the limit of $(f(\mathbf{z}_n))_{n=0}^\infty$ must also exist and be the same for every sequence $(\mathbf{z}_n)_{n=0}^\infty$ in $\mathbb{R}^2$ with $\mathbf{z}_n \rightarrow \mathbf{0}$. (Why?) Consider various choices of $(\mathbf{z}_n)_{n=0}^\infty$.

$$(i) \quad f(x, y) = \frac{\sqrt{xy}}{x^2 + y^2}.$$

$$(ii) \quad f(x, y) = \frac{xy}{x^2 + y^2}.$$

$$(iii) \quad f(x, y) = \frac{xy}{\sqrt{x^2 + y^2}}.$$

- 5.21.†  Prove that unbounded linear transformations are not continuous at the origin (see Definition 3.5.10 for the definition of a bounded linear transformation). Use this to prove that they are not continuous anywhere. Hint: If T is unbounded, construct a sequence of unit vectors $(\mathbf{x}_k)_{k=0}^\infty$, where $\|T(\mathbf{x}_k)\| > k$ for each $k \in \mathbb{N}$. Then modify the sequence so that Theorem 5.3.19 applies.

- 5.22. Let (X, d) be a (not necessarily complete) metric space and assume that $(\mathbf{x}_k)_{k=0}^\infty$ and $(\mathbf{y}_k)_{k=0}^\infty$ are Cauchy sequences. Prove that $(d(\mathbf{x}_k, \mathbf{y}_k))_{k=0}^\infty$ converges.
- 5.23. Which of the following functions are uniformly continuous on the interval $(0, 1) \subset \mathbb{R}$, and which are uniformly continuous on $(0, \infty)$?
- (i) x^3 .
- (ii) $\sin(x)/x$.
- (iii) $x \log(x)$.

Prove that your answer to (i) is correct.

- 5.24.  Prove Proposition 5.4.22. Hint: Prove that a bounded linear transformation is Lipschitz.
- 5.25.  Prove Lemma 5.4.26.
- 5.26. Let $B \subset \mathbb{R}^n$ be a bounded set.

- (i) Let $f : B \rightarrow \mathbb{R}$ be uniformly continuous. Show that the set $f(B)$ is bounded.

- (ii) Give an example to show that this does not necessarily follow if f is merely continuous on B .
- 5.27. Let (X, d) be a metric space such that $d(x, y) < 1$ for all $x, y \in X$, and let $f : X \rightarrow \mathbb{R}$ be uniformly continuous. Does it follow that f must be bounded (that is, that there is an M such that $f(x) < M$ for all $x \in X$)? Justify your answer with either a proof or a counterexample.
- 5.28.* Prove Proposition 5.4.31.

- 5.29. Let $X = (0, 1) \times (0, 1) \subset \mathbb{R}^2$, and let $d(\mathbf{x}, \mathbf{y})$ be the usual Euclidean metric. Give an example of an open cover of X that has no finite subcover.
- 5.30. Let $K \subset \mathbb{R}^n$ be a compact set and $f : K \rightarrow \mathbb{R}^n$ be injective and continuous. Prove that f^{-1} is continuous on $f(K)$.
- 5.31. If $U \subset \mathbb{R}^n$ is open and $K \subset U$ is compact, prove that there is a compact set D such that $K \subset D^\circ$ and $D \subset U$.
- 5.32. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called *periodic* if there exists a number $T > 0$ such that $f(x + T) = f(x)$ for all $x \in \mathbb{R}$. Show that a continuous periodic function is bounded and uniformly continuous on $\mathbb{R}$.
- 5.33.  For any metric space (X, ρ) with $C \subset X$ and $D \subset X$ nonempty subsets, define

$$d(C, D) = \inf\{\rho(\mathbf{c}, \mathbf{d}) \mid \mathbf{c} \in C, \mathbf{d} \in D\}.$$

If C is compact and D is closed, then prove the following:

- (i) $d(C, D) > 0$, whenever C and D are disjoint.
- (ii) If D is also compact, there exists $\mathbf{c}^* \in C$ and $\mathbf{d}^* \in D$ such that $d(C, D) = \rho(\mathbf{c}^*, \mathbf{d}^*)$.
- 5.34.* Prove Proposition 5.5.20.

- 5.35. Prove Proposition 5.6.5.
- 5.36. For each $n \in \mathbb{N}$ let $f_n \in C([0, \pi]; \mathbb{R})$ be given by $f_n(x) = \sin^n(x)$.
- (i) Show that $(f_n)_{n=0}^\infty$ converges pointwise.
- (ii) Describe the function it converges to.
- (iii) Show that $(f_n)_{n=0}^\infty$ does not converge uniformly.
- (iv) Why doesn't this contradict Theorem 5.6.8?
- 5.37. Prove that if $(f_n)_{n=1}^\infty$ and $(g_n)_{n=1}^\infty$ are sequences in $C([a, b]; \mathbb{R})$ that converge uniformly to f and g , respectively, then $(f_n + g_n)_{n=1}^\infty$ converges uniformly to $f + g$.
- 5.38. Prove that if $(f_n)_{n=1}^\infty$ is a sequence in $C([a, b]; \mathbb{R})$ that converges uniformly to f , then for any $a \in \mathbb{R}$ the sequence $(af_n)_{n=1}^\infty$ converges uniformly to af .
- 5.39. Let the sequence $(f_n)_{n=1}^\infty$ in $C([0, 1]; \mathbb{R})$ be given by $f_n(x) = nx/(nx + 1)$. Prove that the sequence converges pointwise, but not uniformly, on $[0, 1]$.
- 5.40. Let the sequence $(f_n)_{n=1}^\infty$ in $C([0, 1]; \mathbb{R})$ be given by $f_n(x) = x/(1 + x^n)$. Prove that the sequence converges pointwise, but not uniformly, on $[0, 1]$.

- 5.41. Show that the sum $\sum_{n=1}^{\infty} (-1)^n/n$ converges, and find a rearrangement of the terms that diverges. Use this to construct an example of a series in $M_2(\mathbb{R})$ that converges, but for which a rearrangement of the terms causes the series to diverge.

- 5.42. Let $(X, \|\cdot\|_X)$ be a Banach space, and let A be a bounded linear operator on X . Prove that the following sums converge absolutely in $\mathcal{B}(X)$:

$$(i) \cos(A) = \sum_{k=0}^{\infty} (-1)^k \frac{A^{2k}}{(2k)!}.$$

$$(ii) \sin(A) = \sum_{k=0}^{\infty} (-1)^k \frac{A^{2k+1}}{(2k+1)!}.$$

$$(iii) \log(A + I) = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{A^k}{k} \text{ if } \|A\| < 1.$$

Hint: Absolute convergence is about sums of real numbers, so you can use the ratio test from introductory calculus.

- 5.43. Prove Proposition 5.7.4. Hint: Proving $A = B^{-1}$ is equivalent to proving that $AB = BA = I$.
- 5.44. Prove Proposition 5.7.8.
- 5.45. Consider the subspace $\mathbb{R}[x] \subset C([-1, 1]; \mathbb{R})$ of polynomials. The Weierstrass approximation theorem (see Volume 2) guarantees that $\mathbb{R}[x]$ is dense in $(C([-1, 1]; \mathbb{R}), \|\cdot\|_{L^\infty})$. Show that the map $D : p(x) \mapsto p'(x)$ is a linear transformation from $\mathbb{R}[x] \rightarrow \mathbb{R}[x] \subset C([-1, 1]; \mathbb{R})$. Show that the function $\sqrt[3]{x}$ is in $C([-1, 1]; \mathbb{R})$ but does not have a derivative in $C([-1, 1]; \mathbb{R})$. This shows that there is no linear extension of D to $(C([-1, 1]; \mathbb{R}), \|\cdot\|_{L^\infty})$. Why doesn't this contradict the continuous linear extension theorem (Theorem 5.7.6)?

- 5.46. Let (X, d) be a metric space. Prove: If the function $f : [0, \infty) \rightarrow [0, \infty)$ is strictly increasing and satisfies $f(0) = 0$ and $f(a + b) \leq f(a) + f(b)$ for all $a, b \in [0, \infty)$, then $\rho(\mathbf{x}, \mathbf{y}) := f(d(\mathbf{x}, \mathbf{y}))$ is a metric on X . Moreover, if f is continuous at 0, then (X, ρ) is topologically equivalent to (X, d) .
- 5.47. Prove that the discrete metric on $\mathbb{F}^n$ is not topologically equivalent to the Euclidean metric. Use this to prove that the discrete metric on $\mathbb{F}^n$ is not induced by any norm.
- 5.48. Prove the claim in the proof of Theorem 5.8.7 that, given any norm $\|\cdot\|$ on a vector space X and any isomorphism $f : X \rightarrow Y$ of vector spaces, the function $\|\cdot\|_f$ on Y defined by $\|\mathbf{y}\|_f = \|f^{-1}(\mathbf{y})\|$ is a norm on Y .
- 5.49. Let X be a set, and let $\mathcal{M}$ be the set of all metrics on X . Prove that topological equivalence is an equivalence relation on $\mathcal{M}$.
- 5.50. Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be equivalent norms on a vector space X , and assume that there exists $0 < \varepsilon < 1$ such that $B_b(0, \varepsilon) \subset B_a(0, 1)$. Prove that

$$\|\mathbf{x}\|_a \leq \frac{\|\mathbf{x}\|_b}{\varepsilon}$$

for all $\mathbf{x} \in B_a(0, 1)$. Hint: If there exists $\mathbf{x} \in B_a(0, 1)$ such that $\|\mathbf{x}\|_a > \|\mathbf{x}\|_b/\varepsilon$, then choose a scalar a so that $\mathbf{y} = a\mathbf{x}$ satisfies $\|\mathbf{y}\|_b/\varepsilon < 1 < \|\mathbf{y}\|_a$. Use this to get a contradiction to the assumption $B_b(0, \varepsilon) \subset B_a(0, 1)$.

-
- 5.51. Prove Proposition 5.9.6.
- 5.52. Prove that there is no homeomorphism from $\mathbb{R}^2$ onto $\mathbb{R}$. Hint: Remove a point from each and consider a connectedness argument.
- 5.53. Consider the metric space $(M_2(\mathbb{R}), d)$, where the metric d is induced by the Frobenius norm, that is, $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_F$. Let X denote the subset of all invertible 2×2 matrices. Is (X, d_X) connected? Prove your answer.
- 5.54. Prove Corollary 5.9.17. Hint: Use the continuous function f to build another continuous function in the style of the proof of the intermediate value theorem.
-
- 5.55. Prove Proposition 5.10.12(ii). Hint: First prove the result for step functions, then for the general case.
- 5.56. Prove Proposition 5.10.12(iv).
- 5.57. Prove Proposition 5.10.12(v). Hint: Use the ε - δ definition of continuity.
- 5.58. Define a function $f(x) : [-1, 1] \rightarrow \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0. \end{cases}$$

It can be shown that this function is Riemann integrable and has Riemann integral equal to 0. Prove that $f(x)$ is not in the space $\overline{S}([-1, 1]; \mathbb{R})$, so the integral of Definition 5.10.7 is not defined for this function.

- 5.59. You cannot always assume that interchanging limits will give the expected results. For example, Exercise 5.45 shows that differentiation and infinite summation do not always commute. In this exercise you show a special case where integration and infinite summation do commute.
- (i) Prove that the integration map $I : \mathbb{R}[x] \rightarrow \mathbb{R}$ given by $I(p) = \int_a^b p(x) dx$ is a bounded linear transformation.
- (ii) Prove that for any polynomial $p(x) = \sum_{k=0}^n c_k x^k$, we have $I(p) = \sum_{k=0}^n c_k (b^{k+1} - a^{k+1}) / (k+1)$.
- (iii) Let $\mathcal{A} \subset C([a, b]; \mathbb{R})$ be the set of absolutely convergent power series on $[a, b]$. Prove that $\mathcal{A}$ is a vector subspace of $C([a, b]; \mathbb{R})$.
- (iv) Prove that the series $\sum_{k=0}^{\infty} c_k (b^{k+1} - a^{k+1}) / (k+1)$ is absolutely convergent if the series $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent on $[a, b]$.
- (v) Let $T : \mathcal{A} \rightarrow \mathbb{R}$ be the map given by termwise integration:

$$T \left(\sum_{k=0}^{\infty} c_k x^k \right) = \sum_{k=0}^{\infty} c_k \int_a^b x^k dx = \sum_{k=0}^{\infty} c_k (b^{k+1} - a^{k+1}) / (k+1).$$

Prove that T is a linear transformation from $\mathcal{A}$ to $\mathbb{R}$.

- (vi) Use the continuous linear extension theorem (Theorem 5.7.6) to prove that $I = T$ on $\mathcal{A}$. In other words, prove that if $f(x) = \sum_{k=0}^{\infty} c_k x^k \in \mathcal{A}$, then

$$I(f) = \int_a^b f(x) dx = \sum_{k=0}^{\infty} c_k (b^{k+1} - a^{k+1}) / (k+1) = T(f).$$

Notes

Some sources for topology include the texts [Mun75, Mor17]. For a brief comparison of the regulated, Riemann, and Lebesgue integrals (for real-valued functions), see [Ber79].

6

Differentiation

In the fall of 1972 President Nixon announced that the rate of increase of inflation was decreasing. This was the first time a sitting president used the third derivative to advance his case for re-election.

—Hugo Rossi

The derivative of a function at a point is a linear transformation describing the best linear approximation to that function in an arbitrarily small neighborhood of that point. This allows us to use many linear analysis tools on nonlinear functions. In single-variable calculus the derivative gives us the slope of the tangent line at a point on the graph, whereas in multidimensional calculus, the derivative gives the individual coordinate slopes of the tangent hyperplane. This generalization of the derivative is ubiquitous in applications. Just as single-variable derivatives are essential to solving problems in single-variable optimization, ordinary differential equations, and univariate probability and statistics, we find that multidimensional derivatives provide the corresponding framework for multivariable optimization problems, partial differential equations, and problems in multivariate probability and statistics.

6.1 The Directional Derivative

We begin by considering differentiability on a vector space. For now we restrict our discussion to functions on $\mathbb{R}^n$, but we revisit these concepts in greater generality in later sections.

6.1.1 Tangent Vectors

Recall the definition of the derivative from single-variable calculus.

Definition 6.1.1. A function $f : (a, b) \rightarrow \mathbb{R}$ is differentiable at $x \in (a, b)$ if the following limit exists:

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}. \quad (6.1)$$

The limit is called the derivative of f at x and is denoted by $f'(x)$. If $f(x)$ is differentiable at every point in (a, b) , we say that f is differentiable on (a, b) .

Remark 6.1.2. The derivative at a point x_0 is important because it defines the best possible linear approximation to f near x_0 . Specifically, the linear transformation $L : \mathbb{R} \rightarrow \mathbb{R}$ given by $L(h) = f'(x_0)h$ provides the best linear approximation of $f(x_0 + h) - f(x_0)$. For every $\varepsilon > 0$ there is a neighborhood $B(0, \delta)$, such that the curve $y = f(x_0 + h) - f(x_0)$ lies between the linear functions $L(h) - \varepsilon h$ and $L(h) + \varepsilon h$, whenever h is in $B(0, \delta)$.

We can easily generalize this definition to a parametrized curve.

Definition 6.1.3. A curve $\gamma : (a, b) \rightarrow \mathbb{R}^n$ is differentiable at $t_0 \in (a, b)$ if the following limit exists:

$$\lim_{h \rightarrow 0} \frac{\gamma(t_0 + h) - \gamma(t_0)}{h}. \quad (6.2)$$

Here the limit is taken with respect to the usual metrics on $\mathbb{R}$ and $\mathbb{R}^n$ (see Definition 5.2.8). If it exists, this limit is called the derivative of $\gamma(t)$ at t_0 and is denoted $\gamma'(t_0)$. If γ is differentiable at every point of (a, b) , we say that γ is differentiable on (a, b) .

Remark 6.1.4. Again, the derivative allows us to define a linear transformation $L(h) = \gamma'(t_0)h$ that is the best linear approximation of γ near t_0 . Said more carefully, $\gamma(t_0 + h) - \gamma(t_0)$ is approximated by $L(h)$ for all h in a sufficiently small neighborhood of 0.

We can write $\gamma(t)$ in standard coordinates as $\gamma(t) = [\gamma_1(t) \ \dots \ \gamma_n(t)]^T$. Each coordinate function $\gamma_i : \mathbb{R} \rightarrow \mathbb{R}$ is a scalar map. Assuming each $\gamma_i(t)$ is differentiable, we can compute the single-variable derivatives $\gamma'_i(t_0)$ using (6.1) at t_0 .

Proposition 6.1.5. A curve $\gamma : (a, b) \rightarrow \mathbb{R}^n$ written in standard coordinates as $\gamma(t) = [\gamma_1(t) \ \dots \ \gamma_n(t)]^T$ is differentiable at t_0 if and only if $\gamma_i(t)$ is differentiable at t_0 for every i . In this case, we have

$$\gamma'(t_0) = [\gamma'_1(t_0) \ \dots \ \gamma'_n(t_0)]^T.$$

Proof. All norms are topologically equivalent in $\mathbb{R}^n$, so we can use any norm to compute the limit. We use the ∞ -norm, which is convenient for this particular proof.

If the derivatives of each coordinate exist, then for any $\varepsilon > 0$ and for each i there exists a δ_i such that whenever $0 < |h| < \delta_i$ we have

$$\left| \frac{\gamma_i(t_0 + h) - \gamma_i(t_0)}{h} - \gamma'_i(t_0) \right| < \varepsilon.$$

Letting $h < \delta = \min(\delta_1, \dots, \delta_n)$ gives

$$\left\| \frac{\gamma(t_0 + h) - \gamma(t_0)}{h} - [\gamma'_1(t_0) \ \dots \ \gamma'_n(t_0)]^T \right\|_{\infty} = \max_i \left| \frac{\gamma_i(t_0 + h) - \gamma_i(t_0)}{h} - \gamma'_i(t_0) \right| < \varepsilon.$$

This shows that $\gamma'(t_0)$ exists and is equal to $[\gamma'_1(t_0) \ \dots \ \gamma'_n(t_0)]^\top$.

Conversely, if $\gamma'(t_0) = [y_1 \ \dots \ y_n]^\top$ exists, then for any i and for any $\varepsilon > 0$ there exists a $\delta > 0$ such that if $0 < |h| < \delta$, then

$$\left| \frac{\gamma_i(t_0 + h) - \gamma_i(t_0)}{h} - y_i \right| \leq \left\| \frac{\gamma(t_0 + h) - \gamma(t_0)}{h} - [y_1 \ \dots \ y_n]^\top \right\|_\infty < \varepsilon,$$

which proves that $\gamma'_i(t_0)$ exists and is equal to y_i . $\square$

Application 6.1.6. If a curve represents the position of a particle as a function of time, then $\gamma'(t_0)$ is the *instantaneous velocity* at t_0 and $\|\gamma'(t_0)\|_2$ is the *speed* at t_0 . We can also take higher-order derivatives; for example, $\gamma''(t_0)$ is the *acceleration* of the particle at time t_0 .

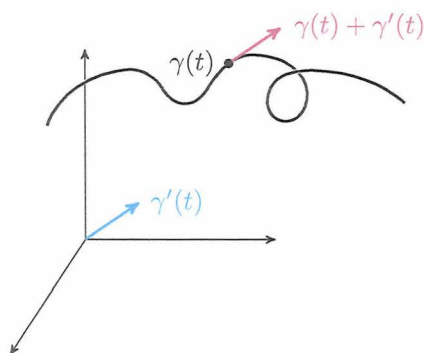


Figure 6.1. The derivative $\gamma'(t)$ of a parametrized curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ points in the direction of the line tangent to the curve at $\gamma(t)$. Note that the tangent vector $\gamma'(t)$ itself (see Definition 6.1.7) is a vector (blue) based at the origin, whereas the line segment (red) from $\gamma(t)$ to $\gamma(t) + \gamma'(t)$ is what is often informally called the “tangent” to the curve.

Definition 6.1.7. Given a differentiable curve $\gamma : (a, b) \rightarrow \mathbb{R}^n$, the tangent vector to the curve γ at t_0 is defined to be $\gamma'(t_0)$. See Figure 6.1 for an illustration.

Example 6.1.8. The curve $\gamma(t) = [\cos t \ \sin t]^\top$ traces out a circle of radius one, centered at the origin. The tangent curve is $\gamma'(t) = [-\sin t \ \cos t]^\top$. The acceleration vector is $\gamma''(t) = [-\cos t \ -\sin t]^\top$ and satisfies $\gamma''(t) = -\gamma(t)$.

Using the product and chain rules for single-variable calculus, we can easily derive the following rules.

Proposition 6.1.9. If the maps $f, g : \mathbb{R} \rightarrow \mathbb{R}^n$ and $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ are differentiable and $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^n$, then

- (i) $(f + g)' = f' + g'$;
- (ii) $(\varphi f)' = \varphi' f + \varphi f'$;
- (iii) $\langle f, g \rangle' = \langle f', g \rangle + \langle f, g' \rangle$;
- (iv) $(f \circ \varphi)'(t) = \varphi'(t)f'(\varphi(t))$.

Proof. The proof follows easily from Proposition 6.1.5 and the standard differentiation rules from single-variable calculus. See Exercise 6.2. $\square$

6.1.2 Directional Derivatives

Another generalization of the single-variable derivative is the directional derivative of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. This is obtained by composing f with a map $\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ given by $\gamma(t) = \mathbf{x} + t\mathbf{u}$ to get a map $f \circ \gamma : \mathbb{R} \rightarrow \mathbb{R}^m$, which can be differentiated as in (6.2). If $\mathbf{u}$ is a unit vector, the derivative of $f(\gamma(t))$ is interpreted to be the derivative of f at $\mathbf{x}$ in the direction of $\mathbf{u}$. If $\mathbf{u}$ is not a unit vector, then the derivative is scaled by a factor of $\|\mathbf{u}\|$. To summarize, we have the following definition.

Definition 6.1.10. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Given $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$, the directional derivative of f at $\mathbf{x}$ with respect to $\mathbf{v}$ is the limit

$$\lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t},$$

if it exists. This limit is often denoted $D_{\mathbf{v}}f(\mathbf{x})$.

Remark 6.1.11. In the next section we prove Theorem 6.2.15, which says that for any fixed $\mathbf{x}$, the function $\phi(\mathbf{v}) = D_{\mathbf{v}}f(\mathbf{x})$ is a linear transformation in $\mathbf{v}$, so $D_{\mathbf{v}_1 + \mathbf{v}_2}f(\mathbf{x}) = D_{\mathbf{v}_1}f(\mathbf{x}) + D_{\mathbf{v}_2}f(\mathbf{x})$. This is an important property of directional derivatives.

Example 6.1.12. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(x, y) = xy^2 + x^3y$. We compute the directional derivative at $\mathbf{x} = (x, y)$ in the direction $\mathbf{v} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ by computing the derivative (with respect to t) of $f(\mathbf{x} + t\mathbf{v})$ at $t = 0$. This gives

$$\begin{aligned} D_{\mathbf{v}}f(x, y) &= \left. \frac{d}{dt} f\left(x + \frac{t}{\sqrt{2}}, y + \frac{t}{\sqrt{2}}\right) \right|_{t=0} \\ &= \left. \frac{d}{dt} \left(\left(x + \frac{t}{\sqrt{2}}\right) \left(y + \frac{t}{\sqrt{2}}\right)^2 + \left(x + \frac{t}{\sqrt{2}}\right)^3 \left(y + \frac{t}{\sqrt{2}}\right) \right) \right|_{t=0} \\ &= \frac{1}{\sqrt{2}}(y^2 + 2xy + 3x^2y + x^3). \end{aligned}$$

6.1.3 Partial Derivatives

Taking directional derivatives along the standard basis vectors $\mathbf{e}_i$ for each i gives what we call *partial derivatives*. In other words, the partial derivatives are the directional derivatives $D_{\mathbf{e}_i}f(\mathbf{x})$, which are often written as $D_i f(\mathbf{x})$ or $\frac{\partial f}{\partial x_i}$.

Definition 6.1.13. Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$. The i th partial derivative of f at the point $\mathbf{x}$ is given by the limit (if it exists)

$$D_i f(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}.$$

Example 6.1.14. In this example we show that the function

$$f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2}, & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0), \end{cases}$$

is not continuous at $(0, 0)$, but its partial derivatives do exist. The sequence $((\frac{1}{n}, \frac{1}{n}))_{n \in \mathbb{N}}$ converges to zero and yet $f(\frac{1}{n}, \frac{1}{n}) = \frac{1}{2}$ for all $n \in \mathbb{N}$; thus f is not continuous at $(0, 0)$. However, we have

$$\begin{aligned} D_1 f(0, 0) &= \lim_{h \rightarrow 0} \frac{f(h, 0) - f(0, 0)}{h} = 0 \quad \text{and} \\ D_2 f(0, 0) &= \lim_{h \rightarrow 0} \frac{f(0, h) - f(0, 0)}{h} = 0. \end{aligned}$$

Thus, the partial derivatives are zero despite the function's failure to be continuous there.

Remark 6.1.15. In the previous definition the i th coordinate is the only one that varies in the limit. Thus, for real-valued functions (that is, when $m = 1$), we can think of this as a single-variable derivative of a scalar function with the only variable being the i th coordinate; the other coordinates can be treated as constants. For vector-valued functions (when $m > 1$), we may compute $D_i f$ using the usual rules for differentiating if we treat x_j as a constant for $j \neq i$.

Example 6.1.16. Consider the function $f(x, y) = xy^2 + x^3y$ of Example 6.1.12. We can compute the first partial derivative of $f(x, y) = xy^2 + x^3y$ by treating f as a function of x only. Thus, the first partial derivative is

$$D_1 f(x, y) = y^2 + 3x^2y.$$

Similarly, we can find the second partial derivative by treating f as a function of y only. Thus, the second partial derivative is

$$D_2 f(x, y) = 2xy + x^3.$$

6.2 The Fréchet Derivative in $\mathbb{R}^n$

In the next two sections we define the derivative of a general function $f : U \rightarrow Y$, where Y is a Banach space and U is an open set in a Banach space. This construction is called the *Fréchet derivative*.

In this section we define the Fréchet derivative for functions on finite-dimensional real vector spaces—this may be familiar to many students.³⁹ In the next section we generalize to arbitrary Banach spaces.

Definition 6.2.1. Let $U \subset \mathbb{R}^n$ be an open set. A function $f : U \rightarrow \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$ if there exists a linear transformation $Df(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \quad (6.3)$$

Here we mean the limit in the sense of Definition 5.2.8. We call $Df(\mathbf{x})$ the derivative of f at $\mathbf{x}$. The derivative is sometimes called the total derivative as a way to distinguish it from the directional and partial derivatives.

Example 6.2.2. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be given by $f(x, y) = (x^3, xy, y^2)$. At the point $\mathbf{x} = (2, 3)$, the function $\mathcal{B}(\mathbb{R}^2; \mathbb{R}^3) \cong M_{3,2}(\mathbb{R})$ given (in standard coordinates) by

$$\begin{bmatrix} 12 & 0 \\ 3 & 2 \\ 0 & 6 \end{bmatrix}$$

is the total derivative of f because

$$\begin{aligned} & \lim_{\mathbf{h} \rightarrow 0} \frac{\left\| f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \begin{bmatrix} 12 & 0 \\ 3 & 2 \\ 0 & 6 \end{bmatrix} \mathbf{h} \right\|}{\|\mathbf{h}\|} \\ &= \lim_{\mathbf{h} \rightarrow 0} \frac{\left\| \begin{bmatrix} (2+h_1)^3 - 2^3 \\ (2+h_1)(3+h_2) - 6 \\ (3+h_2)^2 - 3^2 \end{bmatrix} - \begin{bmatrix} 12 & 0 \\ 3 & 2 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right\|}{\|\mathbf{h}\|} \\ &= \lim_{\mathbf{h} \rightarrow 0} \frac{\left\| \begin{bmatrix} 6h_1^2 + h_1^3 \\ h_1h_2 \\ h_2^2 \end{bmatrix} \right\|}{\|\mathbf{h}\|} = 0. \end{aligned}$$

³⁹Depending on the students' background in multivariable calculus and the amount of class time available, the instructor may wish to skip directly to the general Fréchet derivative in Section 6.3.

It turns out that any norm gives the same answer for the limit, since all norms are topologically equivalent on $\mathbb{R}^n$. Remark 6.3.2 gives more details about different norms and their effect on the derivative.

Example 6.2.3. An argument similar to that in the previous example shows for arbitrary $\mathbf{x} = (x, y)$ that the linear function in $\mathcal{B}(\mathbb{R}^2; \mathbb{R}^3) \cong M_{3,2}$, represented by the matrix

$$\begin{bmatrix} 3x^2 & 0 \\ y & x \\ 0 & 2y \end{bmatrix},$$

is the total derivative of $f(x, y) = (x^3, xy, y^2)$ at $\mathbf{x}$.

Nota Bene 6.2.4. Beware that $Df(\mathbf{x})\mathbf{v}$ is not a linear transformation in $\mathbf{x}$. Rather, for each choice of $\mathbf{x}$, it is a linear transformation in $\mathbf{v}$. So in the previous example, for each fixed $\mathbf{x} = (x, y)$ the transformation

$$Df(\mathbf{x})\mathbf{v} = \begin{bmatrix} 3x^2 & 0 \\ y & x \\ 0 & 2y \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

is linear in $\mathbf{v}$. But

$$Df(a\mathbf{x}) = \begin{bmatrix} 3a^2x^2 & 0 \\ ay & ax \\ 0 & 2ay \end{bmatrix} \neq a \begin{bmatrix} 3x^2 & 0 \\ y & x \\ 0 & 2y \end{bmatrix} = aDf(\mathbf{x}),$$

so $Df : \mathbf{x} \mapsto Df(\mathbf{x})$ is not linear in $\mathbf{x}$.

Example 6.2.5.  If $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear transformation, then for every $\mathbf{x} \in \mathbb{R}^n$ we have

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|L(\mathbf{x} + \mathbf{h}) - L(\mathbf{x}) - L(\mathbf{h})\|}{\|\mathbf{h}\|} = \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{0}\|}{\|\mathbf{h}\|} = 0,$$

so $DL(\mathbf{x})(\mathbf{v}) = L(\mathbf{v})$. Note that $DL(\mathbf{x})$ is independent of $\mathbf{x}$ in this case.

If L is represented in the standard basis by a matrix A , so that $L(\mathbf{x}) = A\mathbf{x}$, then $DL(\mathbf{x})$ is also represented by the matrix A . If the matrix A is equal to $\mathbf{a}^T$, where $\mathbf{a} \in \mathbb{R}^n$, then L can also be written as $L(\mathbf{x}) = \mathbf{x}^T \mathbf{a}$, but the derivative is still $\mathbf{a}^T$. This is an example of a general principle: whenever the independent variable in a function is transposed, then the derivative is also transposed.

Remark 6.2.6. If $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is given by $f(\mathbf{x}) = A\mathbf{x}$ with $A \in M_{m \times n}(\mathbb{R})$, then $Df = A$. Let $(\mathbb{R}^m)^* = \mathcal{B}(\mathbb{R}^m; \mathbb{R})$ be the *dual space* of $\mathbb{R}^m$. We usually write elements of $(\mathbb{R}^m)^*$ as n -dimensional row vectors. A row vector $\mathbf{w}^\top$ corresponds to the function $L \in (\mathbb{R}^m)^*$ given by $L(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$.

The matrix A also defines a function $g : \mathbb{R}^n \rightarrow (\mathbb{R}^m)^*$, given by $g(\mathbf{x}) = \mathbf{x}^\top A$. Exercise 6.10 shows that in this representation the derivative $Dg : \mathbb{R}^n \rightarrow (\mathbb{R}^m)^*$ is the linear transformation $\mathbf{v} \mapsto \mathbf{v}^\top A$.

Alternatively, elements of $(\mathbb{R}^m)^*$ can be represented by column vectors, where $\mathbf{u} \in \mathbb{R}^m$ defines a function $L \in (\mathbb{R}^m)^*$ given by $L(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$. Exercise 6.10 shows that in this form the function g is given by $g(\mathbf{x}) = A^\top \mathbf{x}$, and Dg is the linear transformation $\mathbf{v} \mapsto A^\top \mathbf{v}$.

Remark 6.2.7. If the domain is one dimensional, then a linear transformation $L : \mathbb{R}^1 \rightarrow \mathbb{R}^m$ is given in coordinates by an $m \times 1$ matrix $[\ell_1 \ \dots \ \ell_m]^\top$. In particular, if $m = 1$, then L is just a 1×1 matrix. In traditional single-variable calculus with $f : \mathbb{R} \rightarrow \mathbb{R}$, the derivative $f'(x)$ is a scalar, but we think of $Df(x)$ as an element of $\mathcal{B}(\mathbb{R}; \mathbb{R})$, represented in standard coordinates by the 1×1 matrix $[f'(x)]$.

Example 6.2.8. Suppose $U \subset \mathbb{R}$ and $\gamma : U \rightarrow \mathbb{R}^m$ is a curve. We prove that the derivative $\gamma'(x)$ of the curve defined in Definition 6.1.3 is the total derivative $D\gamma$. Note that

$$\lim_{h \rightarrow 0} \frac{\|\gamma(x+h) - \gamma(x) - \gamma'(x)h\|}{\|h\|} = \lim_{h \rightarrow 0} \left\| \frac{\gamma(x+h) - \gamma(x) - \gamma'(x)h}{h} \right\|.$$

Division by h makes sense because h is a scalar. But this limit is zero if and only if

$$\lim_{h \rightarrow 0} \frac{\gamma(x+h) - \gamma(x)}{h} - \gamma'(x) = 0,$$

which holds if and only if

$$\gamma'(x) = \lim_{h \rightarrow 0} \frac{\gamma(x+h) - \gamma(x)}{h}.$$

Thus, $\gamma'(x) = D\gamma(x)$, as expected.

Remark 6.2.9. The total derivative $Df(\mathbf{x}_0)$, if it exists, defines a linear function that approximates f in a neighborhood of $\mathbf{x}_0$. More precisely, if $L(\mathbf{h}) = Df(\mathbf{x}_0)\mathbf{h}$, then for any $\varepsilon > 0$ there is a $\delta > 0$ such that the function $f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0)$ is within $\varepsilon\|\mathbf{h}\|$ of $L(\mathbf{h})$ (that is, $\|f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - L(\mathbf{h})\| \leq \varepsilon\|\mathbf{h}\|$) whenever $\mathbf{h} \in B(0, \delta)$.

Proposition 6.2.10. Let $U \subset \mathbb{R}^n$ be an open set. If $f : U \rightarrow \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$, then $Df(\mathbf{x})$ is unique.

Proof. Let L_1 and L_2 be two linear transformations satisfying (6.3). For any nonzero $\mathbf{v} \in \mathbb{R}^n$, as $t \rightarrow 0$ we have that

$$\begin{aligned} \frac{\|L_1\mathbf{v} - L_2\mathbf{v}\|}{\|\mathbf{v}\|} &= \frac{\|(f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) - L_2t\mathbf{v}) - (f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) - L_1t\mathbf{v})\|}{|t|\|\mathbf{v}\|} \\ &\leq \frac{\|f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) - L_2t\mathbf{v}\|}{|t|\|\mathbf{v}\|} + \frac{\|f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) - L_1t\mathbf{v}\|}{|t|\|\mathbf{v}\|} \rightarrow 0. \end{aligned}$$

Thus, $L_1\mathbf{v} = L_2\mathbf{v}$ for every $\mathbf{v}$, and hence $L_1 = L_2$. $\square$

Theorem 6.2.11. Let $U \subset \mathbb{R}^n$ be an open set, and let $f : U \rightarrow \mathbb{R}^m$ be given by $f = (f_1, \dots, f_m)$. If f is differentiable on U , then the partial derivatives $D_k f_i(\mathbf{x})$ exist for each $\mathbf{x} \in U$, and the matrix representation of the linear map $Df(\mathbf{x})$ in the standard basis is the Jacobian matrix

$$J(\mathbf{x}) = \begin{bmatrix} D_1 f_1(\mathbf{x}) & D_2 f_1(\mathbf{x}) & \cdots & D_n f_1(\mathbf{x}) \\ D_1 f_2(\mathbf{x}) & D_2 f_2(\mathbf{x}) & \cdots & D_n f_2(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(\mathbf{x}) & D_2 f_m(\mathbf{x}) & \cdots & D_n f_m(\mathbf{x}) \end{bmatrix}. \quad (6.4)$$

Proof. Let $J = [J_1 \ J_2 \ \cdots \ J_n]$ be the matrix representation of $Df(\mathbf{x})$ in the standard basis, with each J_j being a column vector of J . For $\mathbf{h} = h\mathbf{e}_j$ we have

$$\begin{aligned} 0 &= \lim_{h \rightarrow 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} \\ &= \lim_{h \rightarrow 0} \frac{\|f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x}) - hDf(\mathbf{x})\mathbf{e}_j\|}{|h|\|\mathbf{e}_j\|} \\ &= \lim_{h \rightarrow 0} \frac{|f(x_1, \dots, x_j + h, \dots, x_n) - f(x_1, \dots, x_n) - hJ_{ij}|}{|h|}. \end{aligned}$$

Thus, each component also goes to zero as $h \rightarrow 0$, and so

$$\lim_{h \rightarrow 0} \frac{|f_i(x_1, \dots, x_j + h, \dots, x_n) - f_i(x_1, \dots, x_n) - hJ_{ij}|}{|h|} = 0,$$

which implies

$$J_{ij} = \lim_{h \rightarrow 0} \frac{f_i(x_1, \dots, x_j + h, \dots, x_n) - f_i(x_1, \dots, x_n)}{h} = D_j f_i(\mathbf{x}).$$

Thus, (6.4) is the matrix representation of $Df(\mathbf{x})$ in the standard basis. $\square$

Remark 6.2.12. We often denote the Jacobian matrix as $Df(\mathbf{x})$, even though the Jacobian is really only the *matrix representation* of $Df(\mathbf{x})$ in the standard basis.

Example 6.2.13. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be given by

$$f(x, y, z) = (xy + x^2z^2, y^3z^5 + x).$$

The previous proposition shows that the standard representation of $Df(\mathbf{x})$ is given by

$$Df(x, y, z) = \begin{bmatrix} y + 2xz^2 & x & 2x^2z \\ 1 & 3y^2z^5 & 5y^3z^4 \end{bmatrix}.$$

Theorem 6.2.14. Let $U \subset \mathbb{R}^n$ be an open set, and let $f : U \rightarrow \mathbb{R}^m$ be given by $f = (f_1, \dots, f_m)$. If each $D_j f_i(\mathbf{x})$ exists and is continuous on U , then $Df(\mathbf{x})$ exists and is given in standard coordinates by (6.4).

Proof. Let $\mathbf{x} \in U$, and let $J(\mathbf{x})$ be the linear operator with matrix representation in the standard basis given by (6.4). To show that $J(\mathbf{x})$ is $Df(\mathbf{x})$, it suffices to show that (6.3) holds. Since all norms are topologically equivalent, we may use the ∞ -norm. Therefore, it suffices to show that for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every i we have

$$|f_i(\mathbf{y}) - f_i(\mathbf{x}) - [J(\mathbf{x})(\mathbf{y} - \mathbf{x})]_i| < \varepsilon \|\mathbf{y} - \mathbf{x}\|,$$

whenever $0 < \|\mathbf{x} - \mathbf{y}\| < \delta$. Here $[J(\mathbf{x})(\mathbf{x} - \mathbf{y})]_i$ denotes the i th entry of the vector $J(\mathbf{x})(\mathbf{x} - \mathbf{y})$.

For any $\delta > 0$ such that $B(\mathbf{x}, \delta) \subset U$, consider $\mathbf{y} \in B(\mathbf{x}, \delta)$ with $\mathbf{y} \neq \mathbf{x}$. Note that

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= f(y_1, \dots, y_n) - f(x_1, y_2, \dots, y_n) + f(x_1, y_2, \dots, y_n) \\ &\quad - f(x_1, x_2, y_3, \dots, y_n) + f(x_1, x_2, y_3, \dots, y_n) \\ &\quad + \dots + f(x_1, \dots, x_{n-1}, y_n) - f(x_1, \dots, x_n). \end{aligned}$$

For each i, j , let $g_{ij} : \mathbb{R} \rightarrow \mathbb{R}$ be the function $z \mapsto f_i(x_1, \dots, x_{j-1}, z, y_{j+1}, \dots, y_n)$. By the mean value theorem in one dimension, for each i there exists $\xi_{i,1}$ in the interval $[x_1, y_1]$ (or in the interval $[y_1, x_1]$ if $y_1 < x_1$) such that

$$g_{i,1}(y_1) - g_{i,1}(x_1) = D_1 f_i(\xi_{i,1})(y_1 - x_1).$$

Continuing in this manner, we have $\xi_{i,j} \in [x_j, y_j]$ (or in $[y_j, x_j]$ if $y_j < x_j$) such that

$$\begin{aligned} f_i(\mathbf{y}) - f_i(\mathbf{x}) &= D_1 f_i(\xi_{i,1}, y_2, \dots, y_n)(y_1 - x_1) + D_2 f_i(x_1, \xi_{i,2}, y_3, \dots, y_n)(y_2 - x_2) \\ &\quad + \dots + D_n f_i(x_1, x_2, \dots, x_{n-1}, \xi_{i,n})(y_n - x_n) \end{aligned}$$

for every i . Since the i th entry of $J(\mathbf{x})(\mathbf{y} - \mathbf{x})$ is

$$\sum_{j=1}^n D_j f_i(x_1, \dots, x_n)(y_j - x_j),$$

and since $|y_i - x_i| \leq \|\mathbf{y} - \mathbf{x}\|$ we have

$$|f_i(\mathbf{y}) - f_i(\mathbf{x}) - [J(\mathbf{x})(\mathbf{y} - \mathbf{x})]_i| \leq \left(|D_1 f_i(\xi_1, y_2, \dots, y_n) - D_1 f_i(x_1, \dots, x_n)| \right. \\ \left. + \dots + |D_n f_i(x_1, \dots, x_{n-1}, \xi_n) - D_n f_i(x_1, \dots, x_n)| \right) \|\mathbf{y} - \mathbf{x}\|.$$

Since each $D_j f_i(\mathbf{x})$ term is continuous, we can choose δ small enough that

$$|f_i(\mathbf{y}) - f_i(\mathbf{x}) - [J(\mathbf{x})(\mathbf{y} - \mathbf{x})]_i| < \varepsilon \|\mathbf{y} - \mathbf{x}\|,$$

whenever $0 < \|\mathbf{x} - \mathbf{y}\| < \delta$. Hence, $J(\mathbf{x})$ satisfies the definition of the derivative, and $Df(\mathbf{x}) = J(\mathbf{x})$ exists. $\square$

The following theorem shows that the total derivative may be used to compute directional derivatives.

Theorem 6.2.15. *Let $U \subset \mathbb{R}^n$ be an open set. If $f : U \rightarrow \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$, then the directional derivative along $\mathbf{v} \in \mathbb{R}^n$ at $\mathbf{x}$ exists and can be computed as*

$$D_{\mathbf{v}}f(\mathbf{x}) = Df(\mathbf{x})\mathbf{v}. \quad (6.5)$$

In particular, the directional derivative is linear in the direction $\mathbf{v}$.

Proof. Assume $\mathbf{v}$ is nonzero; otherwise the result is trivial. Let

$$\alpha(t) = \left\| \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} - Df(\mathbf{x})\mathbf{v} \right\|. \quad (6.6)$$

It suffices to show that $\lim_{t \rightarrow 0} \alpha(t) = 0$. Choose $\varepsilon > 0$. Since f is differentiable at $\mathbf{x}$, there exists $\delta > 0$ such that

$$\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\| < \varepsilon \|\mathbf{h}\|$$

whenever $0 < \|\mathbf{h}\| < \delta$. Thus, for $\|t\mathbf{v}\| < \delta$, we have

$$\frac{\|f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x}) - tDf(\mathbf{x})\mathbf{v}\|}{|t|} < \varepsilon \|\mathbf{v}\|.$$

But $\|t\mathbf{v}\| < \delta$ if and only if $|t| < \delta \|\mathbf{v}\|^{-1}$; so when $|t| < \delta \|\mathbf{v}\|^{-1}$, we know $\alpha(t) < \varepsilon \|\mathbf{v}\|$. Thus, $\lim_{t \rightarrow 0} \alpha(t) = 0$ and (6.5) holds. $\square$

Example 6.2.16. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(x, y) = xy^2 + x^3y$, as in Example 6.1.12. We have

$$Df(x, y) = [D_1 f(x, y) \quad D_2 f(x, y)] = [y^2 + 3x^2y \quad 2xy + x^3].$$

By Theorem 6.2.15, the directional derivative of f in the direction $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ is

$$D_{(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})} f(x, y) = \begin{bmatrix} y^2 + 3x^2y & 2xy + x^3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \frac{1}{\sqrt{2}}(y^2 + 3x^2y + 2xy + x^3),$$

which agrees with our previous (more laborious) calculation.

Remark 6.2.17. It is possible for the partial derivatives of f to exist even if f is not differentiable. In this case (6.5) may or may not hold; see Exercises 6.7 and 6.8.

6.3 The General Fréchet Derivative

In this section, we extend the idea of the Fréchet derivative from functions on $\mathbb{R}^n$ to functions on general Banach spaces. The Fréchet derivative for functions on arbitrary Banach spaces is very similar to the finite-dimensional case, except that derivatives in the infinite-dimensional case generally have no matrix representations.

Just as the derivative in $\mathbb{R}^n$ can be used for finding local extrema of a function, so also the Fréchet derivative can be used for finding local extrema, but now we consider functions on infinite-dimensional Banach spaces like L^∞ , and the critical values are also elements of that space. This generalization leads to the famous Euler–Lagrange equations of motion in Lagrangian mechanics (see Volume 4).

Throughout this section let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces over $\mathbb{F}$, and let $U \subset X$ be an open set.

6.3.1 The Fréchet Derivative

Definition 6.3.1. A function $f : U \rightarrow Y$ is differentiable at $\mathbf{x} \in U$ if there exists a bounded linear transformation $Df(\mathbf{x}) : X \rightarrow Y$ such that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_Y}{\|\mathbf{h}\|_X} = 0. \quad (6.7)$$

We call $Df(\mathbf{x})$ the derivative of f at $\mathbf{x}$. If f is differentiable at every point $\mathbf{x} \in U$, we say f is differentiable on U .

Remark 6.3.2. Topologically equivalent norms (on either X or Y) give the same derivative. That is, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are two topologically equivalent norms on X and if $\|\cdot\|_r$ and $\|\cdot\|_s$ are two topologically equivalent norms on Y , then Theorem 5.8.4 guarantees that there is an M and an m such that

$$\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_r \leq M\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_s$$

and $\|\mathbf{h}\|_a \geq m\|\mathbf{h}\|_b$. Thus, we have

$$0 \leq \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_r}{\|\mathbf{h}\|_a} \leq \frac{M\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_s}{m\|\mathbf{h}\|_b}.$$

So if f has derivative $Df(\mathbf{x})$ at $\mathbf{x}$ with respect to the norms $\|\cdot\|_s$ and $\|\cdot\|_b$, it must also have the same derivative with respect to $\|\cdot\|_r$ and $\|\cdot\|_a$.

Example 6.3.3.  If $L : X \rightarrow Y$ is any linear transformation, then the argument in Example 6.2.5 shows that $DL(\mathbf{x})(\mathbf{v}) = L(\mathbf{v})$ for every $\mathbf{x}, \mathbf{v} \in X$.

Example 6.3.4. For an example of a linear operator on an infinite-dimensional space, consider $X = C([0, 1]; \mathbb{R})$ with the L^∞ -norm. The function $L : X \rightarrow \mathbb{R}$ given by

$$L(f) = \int_0^1 t f(t) dt$$

is linear in f , so by Example 6.3.3 we have $DL(f)(g) = L(g)$ for every $g \in X$.

Example 6.3.5. If $X = C([0, 1]; \mathbb{R})$ with the L^∞ -norm, then the function $Q : X \rightarrow \mathbb{R}$ given by

$$Q(f) = \int_0^1 t f^3(t) dt$$

is not linear. We show that for each $f \in X$ the derivative $DQ(f)$ is the linear transformation $B : X \rightarrow \mathbb{R}$ defined by $B(g) = \int_0^1 3t f^2(t)g(t) dt$. To see this, compute

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{|Q(f+h) - Q(f) - B(h)|}{\|h\|_{L^\infty}} \\ &= \lim_{h \rightarrow 0} \frac{\left| \int_0^1 t((f+h)^3(t) - f^3(t) - 3f^2(t)h(t)) dt \right|}{\|h\|_{L^\infty}} \\ &= \lim_{h \rightarrow 0} \frac{\left| \int_0^1 t(3f(t)h^2(t) + h^3(t)) dt \right|}{\|h\|_{L^\infty}} \\ &\leq \lim_{h \rightarrow 0} \frac{\|h\|_{L^\infty}^2 \int_0^1 |t(3f(t) + h(t))| dt}{\|h\|_{L^\infty}} = 0. \end{aligned}$$

Definition 6.3.6. Let $f : U \rightarrow Y$ be differentiable on U . If the map $Df : U \rightarrow \mathcal{B}(X; Y)$, given by $\mathbf{x} \mapsto Df(\mathbf{x})$, is continuous, then we say f is continuously differentiable on U . The set of continuously differentiable functions on U is denoted $C^1(U; Y)$.

Proposition 6.3.7. *If $f : U \rightarrow Y$ is differentiable on U , then f is locally Lipschitz at every point⁴⁰ of U ; that is, for all $\mathbf{x}_0 \in U$, there exist $B(\mathbf{x}_0, \delta) \subset U$ and an $L > 0$ such that $\|f(\mathbf{x}) - f(\mathbf{x}_0)\|_Y \leq L\|\mathbf{x} - \mathbf{x}_0\|_X$ whenever $\|\mathbf{x} - \mathbf{x}_0\|_X < \delta$.*

Proof. Let $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$ and $\varepsilon = 1$. Choose $\delta > 0$ so that whenever $0 < \|\mathbf{x} - \mathbf{x}_0\|_X < \delta$, we have

$$\frac{\|f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0)\mathbf{h}\|_Y}{\|\mathbf{x} - \mathbf{x}_0\|_X} < 1,$$

or, alternatively,

$$\|f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|_Y < \|\mathbf{x} - \mathbf{x}_0\|_X. \quad (6.8)$$

Applying the triangle inequality to

$$\|f(\mathbf{x}) - f(\mathbf{x}_0)\|_Y = \|f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|_Y$$

gives

$$\|f(\mathbf{x}) - f(\mathbf{x}_0)\|_Y \leq \|f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|_Y + \|Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|_Y.$$

Combining with (6.8) gives

$$\begin{aligned} \|f(\mathbf{x}) - f(\mathbf{x}_0)\|_Y &\leq \|\mathbf{x} - \mathbf{x}_0\|_X + \|Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\|_Y \\ &\leq (1 + \|Df(\mathbf{x}_0)\|_{X,Y}) \|\mathbf{x} - \mathbf{x}_0\|_X, \end{aligned}$$

where the case of $\mathbf{x} = \mathbf{x}_0$ gives equality. Thus, in the ball $B(\mathbf{x}_0, \delta)$ the function f is locally Lipschitz at $\mathbf{x}_0$ with constant $L = \|Df(\mathbf{x}_0)\|_{X,Y} + 1$. $\square$

Corollary 6.3.8. *If a function is differentiable on an open set U , then it is continuous on U .*

Proof. The proof is Exercise 6.14. $\square$

Remark 6.3.9. Just as in the finite-dimensional case, the derivative is unique. In fact, the proof of Proposition 6.2.10 uses nothing about $\mathbb{R}^n$ and works in the general case exactly as written.

Proposition 6.3.10. *If $f : U \rightarrow Y$ is differentiable at $\mathbf{x} \in U$ and is Lipschitz with constant L in a neighborhood of $\mathbf{x}$, then $\|Df(\mathbf{x})\|_{X,Y} \leq L$.*

Proof. Let $\mathbf{u}$ be a unit vector. For any $\varepsilon > 0$ choose δ such that

$$\frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_Y}{\|\mathbf{h}\|_X} < \varepsilon$$

⁴⁰This should not be confused with the property of being *locally Lipschitz on U* , which is a stronger condition requiring that $\|f(\mathbf{x}) - f(\mathbf{z})\|_Y \leq L\|\mathbf{x} - \mathbf{z}\|_X$ for every $\mathbf{x}, \mathbf{z} \in B(\mathbf{x}_0, \delta)$.

whenever $\|\mathbf{h}\| < \delta$, as in the definition of the derivative. Let $\mathbf{h} = (\delta/2)\mathbf{u}$, which gives

$$\begin{aligned} \|Df(\mathbf{x})\mathbf{u}\|_Y &= \frac{\|Df(\mathbf{x})\mathbf{h}\|_Y}{\|\mathbf{h}\|_X} \\ &\leq \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_Y + \|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})\|_Y}{\|\mathbf{h}\|_X} \\ &\leq \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|_Y}{\|\mathbf{h}\|_X} + \frac{L\|\mathbf{h}\|_X}{\|\mathbf{h}\|_X} \\ &\leq \varepsilon + L. \end{aligned}$$

Since ε was arbitrary, we have the desired result. $\square$

6.3.2 Fréchet Derivatives on Cartesian Products

Recall from Remark 5.8.9 that for $p \in [1, \infty)$ all the p -norms on a Cartesian product of a finite number of Banach spaces are topologically equivalent, and so by Remark 6.3.2 they give the same derivative.

Proposition 6.3.11. *Let $f : U \rightarrow Y_1 \times Y_2 \times \cdots \times Y_m$ be defined by $f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))$, where $((Y_i, \|\cdot\|_{Y_i}))_{i=1}^m$ is a collection of Banach spaces. If f_i is differentiable at $\mathbf{x} \in U$ for each i , then so is f . Moreover, for each $\mathbf{h} \in X$,*

$$Df(\mathbf{x})\mathbf{h} = (Df_1(\mathbf{x})\mathbf{h}, Df_2(\mathbf{x})\mathbf{h}, \dots, Df_m(\mathbf{x})\mathbf{h}).$$

Proof. The proof is similar to that for Proposition 6.1.5 and is Exercise 6.15. $\square$

Definition 6.3.12. *Let $((X_i, \|\cdot\|_i))_{i=1}^n$ be a collection of Banach spaces. Let $f : X_1 \times X_2 \times \cdots \times X_n \rightarrow Y$, where $(Y, \|\cdot\|_Y)$ is a Banach space. The i th partial derivative at $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \in X_1 \times X_2 \times \cdots \times X_n$ is the derivative of the function $g : X_i \rightarrow Y$ defined by $g(\tilde{\mathbf{x}}_i) = f(\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \tilde{\mathbf{x}}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$ and is denoted $D_i f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$.*

Example 6.3.13. In the special case that each X_i is $\mathbb{R}$ and $Y = \mathbb{R}^m$, the definition of partial derivative in Definition 6.3.12 is the same as we gave before in Definition 6.1.13.

Theorem 6.3.14. *Let $((X_i, \|\cdot\|_i))_{i=1}^n$ be a collection of Banach spaces. Let $f : X_1 \times X_2 \times \cdots \times X_n \rightarrow Y$, where $(Y, \|\cdot\|_Y)$ is a Banach space. If f is differentiable at $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, then its partial derivatives $D_i f(\mathbf{x})$ all exist. Moreover, if $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n) \in X_1 \times X_2 \times \cdots \times X_n$, then*

$$Df(\mathbf{x})\mathbf{h} = \sum_{i=1}^n D_i f(\mathbf{x})\mathbf{h}_i.$$

Conversely, if all the partial derivatives of f exist and are continuous on the set $U \subset X_1 \times X_2 \times \cdots \times X_n$, then f is continuously differentiable on U .

Proof. The proof is an easy generalization of Theorems 6.2.14 and 6.2.11. $\square$

6.4 Properties of Derivatives

In this section, we prove three important rules of the derivative, namely, linearity, the product rule, and the chain rule. We also show how to compute the derivative of various matrix-valued functions.

Throughout this section assume that $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces over $\mathbb{F}$ and that $U \subset X$ is an open set.

We begin with a simple observation that simplifies many of the following proofs.

Lemma 6.4.1. *Given a function $f : U \rightarrow Y$, a point $\mathbf{x} \in U$, and a linear transformation $L : X \rightarrow Y$, the following are equivalent:*

- (i) *The function f is differentiable at $\mathbf{x}$ with derivative L .*
- (ii) *For every $\varepsilon > 0$ there is a $\delta > 0$ with $B(\mathbf{x}, \delta) \subset U$ such that*

$$\|f(\mathbf{x} + \xi) - f(\mathbf{x}) - L\xi\|_Y \leq \varepsilon\|\xi\|_X \quad (6.9)$$

whenever $\|\xi\|_X < \delta$.

Proof. When $\xi = \mathbf{0}$ the relation (6.9) is automatically true. When $\xi \neq \mathbf{0}$, dividing by $\|\xi\|_X$ shows that this is equivalent to the limit in the definition of the derivative, except that this inequality is not strict, where the definition of limit has a strict inequality. But this is remedied by choosing a δ such that (6.9) holds with $\varepsilon/2$ instead of ε . $\square$

6.4.1 Linearity

We have already seen that the derivative $Df(\mathbf{x})\mathbf{v}$ is not linear in $\mathbf{x}$, but by definition it is linear in $\mathbf{v}$. The following theorem shows that it is also linear in f .

Theorem 6.4.2 (Linearity). *Assume that $f : U \rightarrow Y$ and $g : U \rightarrow Y$. If f and g are differentiable on U and $a, b \in \mathbb{F}$, then $af + bg$ is also differentiable on U , and $D(af + bg)(\mathbf{x}) = aDf(\mathbf{x}) + bDg(\mathbf{x})$ for each $\mathbf{x} \in U$.*

Proof. Choose $\varepsilon > 0$. Since f and g are differentiable at $\mathbf{x}$, there exists $\delta > 0$ such that $B(\mathbf{x}, \delta) \subset U$ and such that whenever $\|\xi\|_X < \delta$ we have

$$\|f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi\|_Y \leq \frac{\varepsilon\|\xi\|_X}{2(|a| + 1)}$$

and

$$\|g(\mathbf{x} + \xi) - g(\mathbf{x}) - Dg(\mathbf{x})\xi\|_Y \leq \frac{\varepsilon\|\xi\|_X}{2(|b| + 1)}.$$

Thus,

$$\begin{aligned} & \|af(\mathbf{x} + \xi) + bg(\mathbf{x} + \xi) - af(\mathbf{x}) - bg(\mathbf{x}) - aDf(\mathbf{x})\xi - bDg(\mathbf{x})\xi\|_Y \\ & \leq |a|\|f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi\|_Y + |b|\|g(\mathbf{x} + \xi) - g(\mathbf{x}) - Dg(\mathbf{x})\xi\|_Y \\ & \leq \frac{\varepsilon|a|\|\xi\|_X}{2(|a| + 1)} + \frac{\varepsilon|b|\|\xi\|_X}{2(|b| + 1)} \leq \varepsilon\|\xi\|_X \end{aligned}$$

whenever $\|\xi\|_X < \delta$. The result now follows by Lemma 6.4.1. $\square$

Remark 6.4.3. Among other things, the previous proposition tells us that the set $C^1(U; Y)$ of continuously differentiable functions on U is a vector space.

6.4.2 Product Rule

We cannot necessarily multiply vector-valued functions, so a general product rule might not make sense. But if the codomain is a field, then we can talk about the product of functions, and the Fréchet derivative satisfies a product rule.

Theorem 6.4.4 (Product Rule). *If $f : U \rightarrow \mathbb{F}$ and $g : U \rightarrow \mathbb{F}$ are differentiable on U , then the product map $h = fg$ is also differentiable on U , and $Dh(\mathbf{x}) = g(\mathbf{x})Df(\mathbf{x}) + f(\mathbf{x})Dg(\mathbf{x})$ for each $\mathbf{x} \in U$.*

Proof. It suffices to show that

$$\lim_{\xi \rightarrow 0} \frac{|h(\mathbf{x} + \xi) - h(\mathbf{x}) - g(\mathbf{x})Df(\mathbf{x})\xi - f(\mathbf{x})Dg(\mathbf{x})\xi|}{\|\xi\|} = 0. \quad (6.10)$$

Choose $\varepsilon > 0$. Since f and g are differentiable at $\mathbf{x}$, they are locally Lipschitz at $\mathbf{x}$ (see Proposition 6.3.7), so there exists $B(\mathbf{x}, \delta_{\mathbf{x}}) \subset U$ and a constant $L > 0$ such that

$$|f(\mathbf{x} + \xi) - f(\mathbf{x})| \leq L\|\xi\|, \quad (6.11)$$

$$|f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi| \leq \frac{\varepsilon\|\xi\|}{3(\|Df(\mathbf{x})\| + 1)}, \quad (6.12)$$

and

$$|g(\mathbf{x} + \xi) - g(\mathbf{x}) - Dg(\mathbf{x})\xi| \leq \frac{\varepsilon\|\xi\|}{3(\|Dg(\mathbf{x})\| + L)}, \quad (6.13)$$

whenever $\|\xi\| < \delta_{\mathbf{x}}$. If we set $\delta = \min\{\delta_{\mathbf{x}}, \frac{\varepsilon}{3L(\|Dg(\mathbf{x})\| + 1)}, 1\}$, then whenever $\|\xi\| < \delta$, we have that

$$\begin{aligned}
& |f(\mathbf{x} + \xi)g(\mathbf{x} + \xi) - f(\mathbf{x})g(\mathbf{x}) - g(\mathbf{x})Df(\mathbf{x})\xi - f(\mathbf{x})Dg(\mathbf{x})\xi| \\
& \leq |f(\mathbf{x} + \xi)| |g(\mathbf{x} + \xi) - g(\mathbf{x}) - Dg(\mathbf{x})\xi| + \\
& \quad |g(\mathbf{x})| |f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi| + |f(\mathbf{x} + \xi) - f(\mathbf{x})| \|Dg(\mathbf{x})\| \|\xi\| \\
& \leq (|f(\mathbf{x})| + L) |g(\mathbf{x} + \xi) - g(\mathbf{x}) - Dg(\mathbf{x})\xi| + \\
& \quad |g(\mathbf{x})| |f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi| + \delta L \|Dg(\mathbf{x})\| \|\xi\| \\
& \leq \varepsilon \|\xi\|
\end{aligned}$$

whenever $\|\xi\| < \delta$. The result now follows by Lemma 6.4.1. $\square$

Example 6.4.5. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be defined by $f(x, y, z) = x^5y + xy^2 + z^7$, and let $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ be defined by $g(x, y, z) = x^3 + z^{11}$. By the product rule we have

$$\begin{aligned}
D(fg)(x, y, z) &= g(x, y, z)Df(x, y, z) + f(x, y, z)Dg(x, y, z) \\
&= (x^3 + z^{11}) [5x^4y + y^2 \quad x^5 + 2xy \quad 7z^6] + (x^5y + xy^2 + z^7) [3x^2 \quad 0 \quad 11z^{10}].
\end{aligned}$$

For example, the total derivative at the point $(x, y, z) = (0, 1, -1) \in \mathbb{R}^3$ is given by

$$D(fg)(0, 1, -1) = -[1 \quad 0 \quad 7] - [0 \quad 0 \quad 11] = [-1 \quad 0 \quad -18].$$

Proposition 6.4.6. *We have the following differentiation rules:*

- (i) If $\mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x})$ are differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by $f(\mathbf{x}) = \mathbf{u}(\mathbf{x})^\top \mathbf{v}(\mathbf{x})$, then

$$Df(\mathbf{x}) = \mathbf{u}(\mathbf{x})^\top D\mathbf{v}(\mathbf{x}) + \mathbf{v}(\mathbf{x})^\top D\mathbf{u}(\mathbf{x}).$$

- (ii) If $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by $g(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$, then

$$Dg(\mathbf{x}) = \mathbf{x}^\top (A + A^\top).$$

- (iii) Let $\mathbf{w}(\mathbf{x}) = (w_1(\mathbf{x}), \dots, w_m(\mathbf{x}))^\top$ be a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^m$, and let

$$B(\mathbf{x}) = \begin{bmatrix} b_{11}(\mathbf{x}) & b_{12}(\mathbf{x}) & \dots & b_{1m}(\mathbf{x}) \\ b_{21}(\mathbf{x}) & b_{22}(\mathbf{x}) & \dots & b_{2m}(\mathbf{x}) \\ \vdots & \vdots & \dots & \vdots \\ b_{k1}(\mathbf{x}) & b_{k2}(\mathbf{x}) & \dots & b_{km}(\mathbf{x}) \end{bmatrix}$$

be a differentiable function from $\mathbb{R}^n$ to $M_{km}(\mathbb{F})$. If $H : \mathbb{R}^n \rightarrow \mathbb{R}^k$ is given by $H(\mathbf{x}) = B\mathbf{w}$, then

$$DH(\mathbf{x}) = B(\mathbf{x})D\mathbf{w}(\mathbf{x}) + \begin{bmatrix} \mathbf{w}^\top(\mathbf{x})Db_1^\top(\mathbf{x}) \\ \vdots \\ \mathbf{w}^\top(\mathbf{x})Db_k^\top(\mathbf{x}) \end{bmatrix},$$

where $\mathbf{b}_i$ is the i th row of B .

Proof. The proof is Exercise 6.16. $\square$

6.4.3 Chain Rule

The chain rule also holds for Fréchet derivatives.

Theorem 6.4.7 (Chain Rule). *Assume that $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, and $(Z, \|\cdot\|_Z)$ are Banach spaces, that U and V are open neighborhoods of X and Y , respectively, and that $f : U \rightarrow Y$ and $g : V \rightarrow Z$ with $f(U) \subset V$. If f is differentiable on U and g is differentiable on V , then the composite map $h = g \circ f$ is also differentiable on U and $Dh(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x})$ for each $\mathbf{x} \in U$.*

Proof. Let $\mathbf{x} \in U$, and let $\mathbf{y} = f(\mathbf{x})$. Choose $\varepsilon > 0$. Since f is differentiable at $\mathbf{x}$ and locally Lipschitz at $\mathbf{x}$ (by Proposition 6.3.7), there exists $B(\mathbf{x}, \delta_{\mathbf{x}}) \subset U$ and a constant $L > 0$ such that whenever $\|\xi\|_X < \delta_{\mathbf{x}}$ we have

$$\|f(\mathbf{x} + \xi) - f(\mathbf{x}) - Df(\mathbf{x})\xi\|_Y \leq \frac{\varepsilon\|\xi\|_X}{2(\|Dg(\mathbf{y})\|_{Y,Z} + 1)}$$

and

$$\|f(\mathbf{x} + \xi) - f(\mathbf{x})\|_Y \leq L\|\xi\|_X.$$

Since g is differentiable at $\mathbf{y}$, there exists $B(\mathbf{y}, \delta_{\mathbf{y}}) \subset V$ such that whenever $\|\eta\|_Y < \delta_{\mathbf{y}}$ we have

$$\|g(\mathbf{y} + \eta) - g(\mathbf{y}) - Dg(\mathbf{y})\eta\|_Z \leq \frac{\varepsilon\|\eta\|_Y}{2L}.$$

Note that

$$h(\mathbf{x} + \xi) - h(\mathbf{x}) = g(f(\mathbf{x} + \xi)) - g(f(\mathbf{x})) = g(\mathbf{y} + \eta(\xi)) - g(\mathbf{y}),$$

where $\eta(\xi) := f(\mathbf{x} + \xi) - f(\mathbf{x})$. Thus, whenever $\|\xi\|_X < \min\{\delta_{\mathbf{x}}, \delta_{\mathbf{y}}/L\}$, we have that $\|\eta(\xi)\|_Y \leq L\|\xi\|_X < \delta_{\mathbf{y}}$. It follows that

$$\begin{aligned} & \|h(\mathbf{x} + \xi) - h(\mathbf{x}) - Dg(\mathbf{y})Df(\mathbf{x})\xi\|_Z \\ &= \|g(\mathbf{y} + \eta(\xi)) - g(\mathbf{y}) - Dg(\mathbf{y})\eta(\xi) + Dg(\mathbf{y})\eta(\xi) - Dg(\mathbf{y})Df(\mathbf{x})\xi\|_Z \\ &\leq \|g(\mathbf{y} + \eta(\xi)) - g(\mathbf{y}) - Dg(\mathbf{y})\eta(\xi)\|_Z + \|Dg(\mathbf{y})\|_{Y,Z}\|\eta(\xi) - Df(\mathbf{x})\xi\|_Y \\ &\leq \frac{\varepsilon\|\eta(\xi)\|_Y}{2L} + \|Dg(\mathbf{y})\|_{Y,Z} \frac{\varepsilon\|\xi\|_X}{2(\|Dg(\mathbf{y})\|_{Y,Z} + 1)} \\ &\leq \varepsilon\|\xi\|_X. \end{aligned}$$

The result now follows by Lemma 6.4.1. $\square$

Example 6.4.8. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ both be differentiable. If $g(p, q) = (x(p, q), y(p, q))$ and $h(p, q) = f(g(p, q)) = f(x(p, q), y(p, q))$, then

$$Dg(p, q) = \begin{bmatrix} \frac{\partial x}{\partial p} & \frac{\partial x}{\partial q} \\ \frac{\partial y}{\partial p} & \frac{\partial y}{\partial q} \end{bmatrix} \quad \text{and} \quad Df(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix},$$

and by the chain rule we have $Dh(p, q) = Df(x, y)Dg(p, q)$. Hence,

$$\begin{bmatrix} \frac{\partial h}{\partial p} & \frac{\partial h}{\partial q} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial p} & \frac{\partial x}{\partial q} \\ \frac{\partial y}{\partial p} & \frac{\partial y}{\partial q} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \frac{\partial x}{\partial p} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial p} & \frac{\partial f}{\partial x} \frac{\partial x}{\partial q} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial q} \end{bmatrix}.$$

Example 6.4.9. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be arbitrary and $\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ be defined by $\gamma(t) = \mathbf{x} + t\mathbf{v}$. The chain rule gives

$$D_{\mathbf{v}}f(\mathbf{x}) = \left. \frac{d}{dt}f(\gamma(t)) \right|_{t=0} = Df(\gamma(0))\gamma'(0) = Df(\mathbf{x})\mathbf{v}.$$

In other words, the directional derivative $D_{\mathbf{v}}f(\mathbf{x})$ is the product of the derivative $Df(\mathbf{x})$ and the tangent vector $\mathbf{v}$. This is an alternative proof of Theorem 6.2.15.

6.5 Mean Value Theorem and Fundamental Theorem of Calculus

In this section, we generalize several more properties of single-variable derivatives to functions on Banach spaces. These include both the mean value theorem and the fundamental theorem of calculus. We then use these to describe some situations where uniform limits commute with differentiation.

Throughout this section, unless indicated otherwise, assume that $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces over $\mathbb{F}$ and that $U \subset X$ is an open set.

6.5.1 The Mean Value Theorem

We begin this section by generalizing the mean value theorem to functions on a Banach space.

Theorem 6.5.1 (Mean Value Theorem). *Let $f : U \rightarrow \mathbb{R}$ be differentiable on U . Given $\mathbf{a}, \mathbf{b} \in U$, if the entire line segment $\ell(\mathbf{a}, \mathbf{b}) = \{(1-t)\mathbf{a} + t\mathbf{b} \mid t \in (0, 1)\}$ is also in U , then there exists $\mathbf{c} \in \ell(\mathbf{a}, \mathbf{b})$ such that*

$$f(\mathbf{b}) - f(\mathbf{a}) = Df(\mathbf{c})(\mathbf{b} - \mathbf{a}). \quad (6.14)$$

Proof. Let $h : [0, 1] \rightarrow \mathbb{R}$ be given by $h(t) = f((1-t)\mathbf{a} + t\mathbf{b})$. Since f is differentiable on U and U contains the line segment $\ell(\mathbf{a}, \mathbf{b})$, the function h is differentiable on $(0, 1)$ and continuous on $[0, 1]$. By the usual mean value theorem in one dimension, there exists $t_0 \in (0, 1)$ such that $h(1) - h(0) = h'(t_0)$. By the chain rule, $h'(t_0) = Df((1-t_0)\mathbf{a} + t_0\mathbf{b})(\mathbf{b} - \mathbf{a}) = Df(\mathbf{c})(\mathbf{b} - \mathbf{a})$, and therefore (6.14) holds. $\square$

Remark 6.5.2. There is an important property in mathematics that allows the hypothesis in Theorem 6.5.1 to be stated more concisely. If for *any* two points $\mathbf{a}$

and $\mathbf{b}$ in the set U the line segment $\ell(\mathbf{a}, \mathbf{b})$ is also contained in U , then we say that U is *convex*. See Figure 6.2 for an illustration. Convexity is a property that is used widely in applications. We treat convexity in much more depth in Volume 2.



Figure 6.2. In this figure the set V is not convex because the line segment $\ell(\mathbf{a}, \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$ does not lie inside of V . But the set U is convex because for every pair of points $\mathbf{a}, \mathbf{b} \in U$ the line segment $\ell(\mathbf{a}, \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$ lies in U ; see Remark 6.5.2.

6.5.2 Single-Variable Fundamental Theorem of Calculus

Lemma 6.5.3. Assume $f : [a, b] \rightarrow X$ is continuous on $[a, b]$ and differentiable on (a, b) . If $Df(t) = 0$ for all $t \in (a, b)$, then f is constant.

Proof. Let $\alpha, \beta \in (a, b)$ with $\alpha < \beta$. Given $\varepsilon > 0$ and $t \in (\alpha, \beta)$, there exists $\delta_t > 0$ such that $\|f(t+h) - f(t)\|_X \leq \varepsilon|h|$ whenever $|h| < \delta_t$. Since $[\alpha, \beta]$ is compact, we can cover it with a finite set of overlapping intervals $\{(t_i - \delta_i, t_i + \delta_i)\}_{i=1}^n$, where, without loss of generality, we can assume $\alpha < t_1 < t_2 < \dots < t_n < \beta$. Choose points $x_0, x_1, \dots, x_n$ so that

$$\alpha = x_0 < t_1 < x_1 < t_2 < \dots < t_n < x_n = \beta,$$

where $|x_i - t_i| < \delta_i$ and $|x_i - t_{i-1}| < \delta_{i-1}$ for each i ; that is, the x_i are chosen to be in the overlapping regions created by adjacent open intervals. Thus, we have

$$\begin{aligned} \|f(\beta) - f(\alpha)\|_X &= \left\| \sum_{i=1}^n [f(x_i) - f(x_{i-1})] \right\|_X \\ &= \left\| \sum_{i=1}^n [(f(x_i) - f(t_i)) + (f(t_i) - f(x_{i-1}))] \right\|_X \\ &\leq \sum_{i=1}^n [\|f(x_i) - f(t_i)\|_X + \|f(t_i) - f(x_{i-1})\|_X] \\ &\leq \sum_{i=1}^n [\varepsilon(x_i - t_i) + \varepsilon(t_i - x_{i-1})] \\ &= \varepsilon(\beta - \alpha). \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, as are $\alpha, \beta \in (a, b)$, it follows that f is constant on (a, b) . Since f is continuous on $[a, b]$, it must also be constant on $[a, b]$. $\square$

Theorem 6.5.4 (Fundamental Theorem of Calculus). *Using the natural isomorphism $\mathcal{B}(\mathbb{R}; X) \cong X$ defined by sending $\phi \in \mathcal{B}(\mathbb{R}; X)$ to $\phi(1) \in X$, we have the following:*

(i) *If $f \in C([a, b]; X)$, then for all $t \in (a, b)$ we have*

$$\frac{d}{dt} \int_a^t f(s) ds = f(t). \quad (6.15)$$

(ii) *If $F : [a, b] \rightarrow X$ is continuously differentiable on (a, b) and $DF(t)$ extends to a continuous function on $[a, b]$, then*

$$\int_a^b DF(s) ds = F(b) - F(a). \quad (6.16)$$

Proof.

(i) Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that $\|f(t+h) - f(t)\|_X < \varepsilon$ whenever $|h| < \delta$. Thus,

$$\begin{aligned} \left\| \int_a^{t+h} f(s) ds - \int_a^t f(s) ds - f(t)h \right\|_X &= \left\| \int_t^{t+h} f(s) - f(t) ds \right\|_X \\ &\leq \left| \int_t^{t+h} \|f(s) - f(t)\|_X ds \right| \leq |h|\eta(h), \end{aligned}$$

where $\eta(h) = \sup_{s \in B(t, |h|)} \|f(s) - f(t)\|_X < \varepsilon$. Hence, (6.15) holds.

(ii) Let $G(t) = \int_a^t DF(s) ds - F(t)$. This is continuous by Proposition 5.10.12(v). Moreover, by (i) above, we have that $DG(t) = 0$ for all $t \in (a, b)$. Therefore, by Lemma 6.5.3, we have that $G(t)$ is constant. In particular, $G(a) = G(b)$, which implies (6.16). $\square$

Corollary 6.5.5 (Integral Mean Value Theorem). *Let $f \in C^1(U; Y)$. If the line segment $\ell(\mathbf{x}_*, \mathbf{x}) = \{(1-t)\mathbf{x}_* + t\mathbf{x} \mid t \in [0, 1]\}$ is contained in U , then*

$$f(\mathbf{x}) - f(\mathbf{x}_*) = \int_0^1 Df(t\mathbf{x} + (1-t)\mathbf{x}_*)(\mathbf{x} - \mathbf{x}_*) dt.$$

Alternatively, if we let $\mathbf{x} = \mathbf{x}_ + \mathbf{h}$, then*

$$f(\mathbf{x}_* + \mathbf{h}) - f(\mathbf{x}_*) = \int_0^1 Df(\mathbf{x}_* + t\mathbf{h})\mathbf{h} dt. \quad (6.17)$$

Moreover, we have

$$\|f(\mathbf{x}) - f(\mathbf{x}_*)\|_Y \leq \sup_{\mathbf{c} \in \ell(\mathbf{x}_*, \mathbf{x})} \|Df(\mathbf{c})\|_{X,Y} \|\mathbf{x} - \mathbf{x}_*\|_X. \quad (6.18)$$

Proof. The proof is Exercise 6.22. $\square$

Corollary 6.5.6 (Change of Variable Formula). *Let $f \in C([a, b]; X)$ and $g : [c, d] \rightarrow [a, b]$ be continuous. If g' is continuous on (c, d) and can be continuously extended to $[c, d]$, then*

$$\int_c^d f(g(s))g'(s) ds = \int_{g(c)}^{g(d)} f(\tau) d\tau. \quad (6.19)$$

Proof. The proof is Exercise 6.23. $\square$

6.5.3 Uniform Convergence and Derivatives

Theorem 5.7.5 shows that for any Banach space X , the integral is a bounded linear operator, and hence continuous, on $\overline{S([a, b]; X)}$ in $L^\infty([a, b]; X)$. This means that for any uniformly convergent sequence $(f_n)_{n=0}^\infty$ in $\overline{S([a, b]; X)}$, the integral commutes with the limit; that is,

$$\lim_{n \rightarrow \infty} \int_a^b f_n dt = \int_a^b \left(\lim_{n \rightarrow \infty} f_n \right) dt.$$

It is natural to hope that something similar would be true for derivatives. Unfortunately, the derivative is not a bounded linear transformation, which means, among other things, that the derivatives of a convergent sequence may not even converge. But all is not lost. The mean value theorem allows us to prove that derivatives do pass through limits, provided the necessary limit exists.

First, however, we must be careful about what we mean by uniform convergence for differentiable functions. The problem is that differentiability is a property of functions on open sets, but the L^∞ -norm is best behaved on compact sets. Specifically, a function $f \in C^1(U; Y)$ does not generally have a finite L^∞ -norm (for example, $f(x) = 1/x$ on the open set $(0, 1)$ is differentiable, but has $\|f\|_{L^\infty} = \infty$). The solution, when studying convergence of functions in the L^∞ -norm, is to look at all the compact subsets of U .

Definition 6.5.7. *Restricting a function $f \in C(U; Y)$ to a compact subset $K \subset U$ defines a function $f|_K \in (C(K; Y), \|\cdot\|_{L^\infty})$, where $\|f|_K\|_{L^\infty} = \sup_{\mathbf{x} \in K} \|f(\mathbf{x})\|$. We say that a sequence $(f_n)_{n=0}^\infty \in C(U; Y)$*

- (i) *is a Cauchy sequence in $C(U; Y)$ if the restriction $(f_n|_K)_{n=0}^\infty \in (C(K; Y), \|\cdot\|_{L^\infty})$ is a Cauchy sequence for every compact subset $K \subset U$;*
- (ii) *converges uniformly on compact subsets to $f \in C(U; Y)$ if the restriction $(f_n|_K)_{n=0}^\infty$ converges to $f|_K$ in $(C(K; Y), \|\cdot\|_{L^\infty})$ for every compact subset $K \subset U$.*

Remark 6.5.8. To check that a sequence on an open subset U of a finite-dimensional Banach space is Cauchy or that it converges uniformly on compact subsets, it suffices to check the condition on compact subsets that are closed balls of the form $\overline{B(\mathbf{x}, r)}$. Moreover, when the open set U is of the form $B(\mathbf{x}_0, R)$, then it suffices to check the condition on closed balls $\overline{B(\mathbf{x}_0, r)}$ centered at $\mathbf{x}_0$. The proof of this is Exercise 6.24.

Example 6.5.9. For each $n \in \mathbb{N}$, let $f_n(x) = x^n$. Each f_n is differentiable on the interval $(0, 1)$ and has $\sup_{(0,1)} f_n = 1$ for all n . But any compact subset of $(0, 1)$ lies in an interval $[a, b]$ with $0 < a < b < 1$, and we have $\|x^n\|_{[a,b]} = b^n \rightarrow 0$ as $n \rightarrow \infty$, so $f_n \rightarrow 0$ uniformly on compact subsets in the open interval $(0, 1)$.

Unexample 6.5.10. Consider the sequence $g_n = \cos(nx)/n$ on the open set $U = (0, 2\pi)$. This sequence converges uniformly on compact subsets to 0, but the sequence of derivatives $g'_n = -\sin(nx)$ does not converge at all. This example also shows that the induced norm of the derivative as a linear operator is infinite, since for every $n > 0$ we have

$$\left\| \frac{d}{dx} \right\| = \sup_f \frac{\left\| \frac{df}{dx} \right\|_{L^\infty}}{\|f\|_{L^\infty}} \geq \frac{\left\| \frac{d}{dx} g_n \right\|_{L^\infty}}{\|g_n\|_{L^\infty}} = \frac{1}{1/n} = n.$$

Nevertheless, we can prove the following important result about uniform convergence of derivatives in a finite-dimensional space.

Theorem 6.5.11. *Let X be finite-dimensional Banach space. Fix an open ball $U = B_X(\mathbf{x}_*, r) \subset X$ with a sequence $(f_n)_{n=0}^\infty \subset C^1(U; Y)$ such that $(f_n(\mathbf{x}_*))_{n=0}^\infty \subset Y$ converges. If $(Df_n)_{n=0}^\infty$ converges uniformly on compact subsets to $g \in C(U; \mathcal{B}(X; Y))$, then the sequence $(f_n)_{n=0}^\infty$ converges uniformly on compact subsets to a function $f \in C^1(U; Y)$, and $g = Df$.*

The idea of the proof is fairly simple—define the limit function using the integral mean value theorem, and then use the fact that uniform limits commute with integration. The result can also be extended to any path-connected U .

Proof. Let $\mathbf{z} = \lim_{n \rightarrow \infty} f_n(\mathbf{x}_*)$. For each $\mathbf{x} \in U$ let $\mathbf{h} = \mathbf{x} - \mathbf{x}_*$, and define

$$f(\mathbf{x}) = \mathbf{z} + \int_0^1 g(\mathbf{x}_* + t\mathbf{h})\mathbf{h} dt. \quad (6.20)$$

This makes sense because U is convex. Note that $f(\mathbf{x}_*) = \mathbf{z} = \lim_{n \rightarrow \infty} f_n(\mathbf{x}_*)$.

We claim that $(f_n)_{n=0}^\infty$ converges to f uniformly on compact subsets. To prove this, it suffices (by Remark 6.5.8) to prove uniform convergence on any compact ball $K = \overline{B}(\mathbf{x}_*, \rho) \subset U$. To prove convergence on such a K , note that for any $\varepsilon > 0$ and for each $\mathbf{x} \in K$ we have, by the integral mean value theorem,

$$\begin{aligned} \|f_n(\mathbf{x}) - f(\mathbf{x})\|_Y &= \left\| f_n(\mathbf{x}_*) + \int_0^1 Df_n(\mathbf{x}_* + t\mathbf{h})\mathbf{h} dt - \mathbf{z} - \int_0^1 g(\mathbf{x}_* + t\mathbf{h})\mathbf{h} dt \right\|_Y \\ &\leq \|f_n(\mathbf{x}_*) - \mathbf{z}\|_Y + \left\| \int_0^1 (Df_n(\mathbf{x}_* + t\mathbf{h}) - g(\mathbf{x}_* + t\mathbf{h}))\mathbf{h} dt \right\|_Y \\ &\leq \|f_n(\mathbf{x}_*) - \mathbf{z}\|_Y + \sup_{\mathbf{c} \in K} \|Df_n(\mathbf{c}) - g(\mathbf{c})\|_{X,Y} \|\mathbf{h}\|_X. \end{aligned} \quad (6.21)$$

Because $f_n(\mathbf{x}_*) \rightarrow \mathbf{z}$, there is an $N > 0$ such that $\|f_n(\mathbf{x}_*) - \mathbf{z}\|_Y < \varepsilon/2$ if $n \geq N$, and since $Df_n \rightarrow g$ uniformly on K , there is an $M > N$ such that

$$\sup_{\mathbf{c} \in K} \|Df_n(\mathbf{c}) - g(\mathbf{c})\|_{X,Y} < \frac{\varepsilon}{2r}$$

whenever $n > M$. Combined with the fact that $\|\mathbf{h}\|_X < r$, Equation (6.21) gives

$$\|f_n(\mathbf{x}) - f(\mathbf{x})\|_Y < \frac{\varepsilon}{2} + \frac{\varepsilon}{2r} \|\mathbf{h}\|_X < \frac{\varepsilon}{2} + \frac{\varepsilon}{2r} r = \varepsilon$$

whenever $n \geq M$. Since the choice of N and M was independent of $\mathbf{x}$, we have $\|f_n - f\|_{L^\infty} \leq \varepsilon$ on K . Therefore, $f_n \rightarrow f$ uniformly on all closed balls $\overline{B(\mathbf{x}_*, \rho)} \subset U$, and therefore on all compact subsets of U .

Finally, we must show that $Df(\mathbf{x}) = g(\mathbf{x})$ for any $\mathbf{x}$ in U . For any $\varepsilon > 0$ we must find a δ such that $\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - g(\mathbf{x})\mathbf{h}\|_Y \leq \varepsilon \|\mathbf{h}\|_X$, whenever $\|\mathbf{h}\|_X < \delta$. By the integral mean value theorem, we have

$$f_n(\mathbf{x} + \mathbf{h}) - f_n(\mathbf{x}) = \int_0^1 Df_n(\mathbf{x} + t\mathbf{h})\mathbf{h} dt.$$

Because U is open, there is an $a > 0$ such that $\overline{B(\mathbf{x}, a)} \subset U$. Integration is a bounded linear operator, so it commutes with uniform limits. Therefore, for all $\|\mathbf{h}\| < a$ we have $f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = \int_0^1 g(\mathbf{x} + t\mathbf{h})\mathbf{h} dt$, and thus

$$\begin{aligned} \|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - g(\mathbf{x})\mathbf{h}\|_Y &= \left\| \int_0^1 g(\mathbf{x} + t\mathbf{h})\mathbf{h} dt - g(\mathbf{x})\mathbf{h} \right\|_Y \\ &= \left\| \int_0^1 (g(\mathbf{x} + t\mathbf{h}) - g(\mathbf{x}))\mathbf{h} dt \right\|_Y. \end{aligned}$$

Continuity of g on the compact set $\overline{B(\mathbf{x}, a)}$ implies g is uniformly continuous there, and hence there exists a δ with $0 < \delta < a$ such that $\|g(\mathbf{x} + t\mathbf{h}) - g(\mathbf{x})\|_{X,Y} < \varepsilon$ whenever $\|\mathbf{h}\|_X < \delta$. Therefore,

$$\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - g(\mathbf{x})\mathbf{h}\|_Y \leq \varepsilon \|\mathbf{h}\|_X,$$

as required. $\square$

6.6 Taylor's Theorem

Taylor's Theorem is one of the most powerful tools in analysis. It allows us to approximate smooth (differentiable) functions in a small neighborhood to arbitrary precision using polynomials. Not only does this give us a lot of insight into the behavior of the function, but it also often allows us to compute, or at least approximate, functions that are otherwise difficult to compute. It is also the foundation of many of the important theorems of applied mathematics.

Before we can describe Taylor's theorem, we need to discuss higher-order derivatives. Throughout this section, let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces over $\mathbb{F}$, and let $U \subset X$ be an open set.

6.6.1 Higher-Order Derivatives

Definition 6.6.1. For $k \geq 2$, let $\mathcal{B}^k(X; Y)$ be defined inductively as $\mathcal{B}^k(X; Y) = \mathcal{B}(X; \mathcal{B}^{k-1}(X; Y))$ and $\mathcal{B}^1(X; Y) = \mathcal{B}(X; Y)$.

Definition 6.6.2. Let $f : U \rightarrow Y$ be differentiable on U . If $Df : U \rightarrow \mathcal{B}(X; Y)$ is differentiable, then we denote the derivative of Df as $D^2f(\mathbf{x}) \in \mathcal{B}(X; \mathcal{B}(X; Y))$, or equivalently as $D^2f(\mathbf{x}) \in \mathcal{B}^2(X; Y)$, and call this map the second derivative.

Proceeding inductively, if the map $D^{k-1}f : U \rightarrow \mathcal{B}^{k-1}(X; Y)$ is differentiable for $k > 2$, then we denote the k th derivative as $D^kf(\mathbf{x}) \in \mathcal{B}^k(X; Y)$. If the k th derivative $D^kf(\mathbf{x})$ is continuous on U , we say that f is k -times continuously differentiable on U and denote the space of such functions as $C^k(U; Y)$.

Finally, we say that f is smooth and write $f \in C^\infty(U; Y)$ if $f \in C^k(U; Y)$ for every positive integer k .

Example 6.6.3.  If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, then for each $\mathbf{x} \in \mathbb{R}^n$ we have $Df(\mathbf{x}) \in \mathcal{B}(\mathbb{R}^n; \mathbb{R})$. By the Riesz representation theorem (Theorem 3.7.1), we have $\mathcal{B}(\mathbb{R}^n; \mathbb{R}) \cong \mathbb{R}^n$, where a vector $\mathbf{u} \in \mathbb{R}^n$ corresponds to the function $\mathbf{v} \mapsto \langle \mathbf{u}, \mathbf{v} \rangle$ of $\mathcal{B}(\mathbb{R}^n; \mathbb{R})$. In the standard basis on $\mathbb{R}^n$, it is convenient to write an element of $\mathcal{B}(\mathbb{R}^n; \mathbb{R})$ as a row vector $\mathbf{u}^\top$, so that the corresponding linear transformation is just given by matrix multiplication (to indicate this we write $\mathcal{B}^1(\mathbb{R}^n; \mathbb{R}) \cong (\mathbb{R}^n)^\top$). This is what is meant when Theorem 6.2.11 says that (in the standard basis) $Df(\mathbf{x})$ can be written as a row vector $Df(\mathbf{x}) = [D_1f(\mathbf{x}) \ \cdots \ D_nf(\mathbf{x})]$.

If $Df : \mathbb{R}^n \rightarrow \mathcal{B}(\mathbb{R}^n; \mathbb{R}) \cong (\mathbb{R}^n)^\top$ is also differentiable at each $\mathbf{x} \in \mathbb{R}^n$, then $D^2f(\mathbf{x}) \in \mathcal{B}(\mathbb{R}^n; \mathcal{B}(\mathbb{R}^n; \mathbb{R})) \cong \mathcal{B}(\mathbb{R}^n; (\mathbb{R}^n)^\top)$. Theorem 6.2.11 still applies, but since $Df(\mathbf{x})$ is a row vector, the second derivative in the $\mathbf{u}$ direction $D^2f(\mathbf{x})(\mathbf{u}) \in \mathcal{B}^1(\mathbb{R}^n; \mathbb{R})$ is also a row vector. In the standard basis we have $D^2f(\mathbf{x})(\mathbf{u}) = \mathbf{u}^\top H^\top \in \mathcal{B}^1(\mathbb{R}^n; \mathbb{R}) \cong (\mathbb{R}^n)^\top$ and $D^2f(\mathbf{x})(\mathbf{u})(\mathbf{v}) = \mathbf{u}^\top H^\top \mathbf{v}$, where

$$\begin{aligned} H &= (D [D_1f(\mathbf{x}) \ \cdots \ D_nf(\mathbf{x})]^\top) \\ &= \begin{bmatrix} D_1D_1f(\mathbf{x}) & \cdots & D_nD_1f(\mathbf{x}) \\ D_1D_2f(\mathbf{x}) & \cdots & D_nD_2f(\mathbf{x}) \\ \vdots & & \vdots \\ D_1D_nf(\mathbf{x}) & \cdots & D_nD_nf(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}. \end{aligned}$$

The matrix H is called the *Hessian* of f . The next proposition shows that H is symmetric.

Definition 6.6.4. Let $((X_i, \|\cdot\|_i))_{i=1}^n$ be a collection of Banach spaces. Fix an open set $U \subset X_1 \times \cdots \times X_n$ and an ordered list of k integers $i_1, \dots, i_k$ between 1 and n (not necessarily distinct). The k th-order partial derivative of $f \in C^k(U; Y)$ corresponding to $i_1, \dots, i_k$ is the function $D_{i_1}D_{i_2} \cdots D_{i_k}f$ (see Definition 6.3.12).

If $Y = \mathbb{F}$ and $X_i = \mathbb{F}$ for every i , then $U \subset \mathbb{F}^n$, and in this case we often write $D_{i_1}D_{i_2}\cdots D_{i_k}f$ as

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} = D_{i_1}D_{i_2}\cdots D_{i_k}f.$$

Proposition 6.6.5. *Let $f \in C^2(U; Y)$, where Y is finite dimensional. For any $\mathbf{x} \in U$ and any $\mathbf{v}, \mathbf{w}$ in X , we have*

$$D^2f(\mathbf{x})(\mathbf{v}, \mathbf{w}) = D^2f(\mathbf{x})(\mathbf{w}, \mathbf{v}). \quad (6.22)$$

If $X = X_1 \times \cdots \times X_n$, then this says

$$D_i D_j f(\mathbf{x}) = D_j D_i f(\mathbf{x}) \quad (6.23)$$

for all i and j . In the case that $X = \mathbb{F} \times \cdots \times \mathbb{F} = \mathbb{F}^n$ and $Y = \mathbb{F}^m$ with $f: X \rightarrow Y$ given by $f = (f_1, \dots, f_m)$, then this is equivalent to

$$\frac{\partial^2 f_k}{\partial x_i \partial x_j} = \frac{\partial^2 f_k}{\partial x_j \partial x_i} \quad (6.24)$$

for all i, j , and k .

Proof. Since Y is finite dimensional, we may assume that $Y \cong \mathbb{F}^m$ and that $f = (f_1, \dots, f_m)$. The usual norm on $\mathbb{C}^m$ is the same as the usual norm on $\mathbb{R}^{2m}$, via the standard map $(x_1 + iy_1, \dots, x_m + iy_m) \mapsto (x_1, y_1, \dots, x_m, y_m)$; therefore, we may assume that $\mathbb{F} = \mathbb{R}$. Moreover, it suffices to prove the theorem for each f_k individually, so we may assume that $Y = \mathbb{R}^1$.

For each $\mathbf{x} \in U$ let $g_t(\mathbf{x}) = f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})$, and let $S_{s,t}(\mathbf{x}) = g_t(\mathbf{x} + s\mathbf{w}) - g_t(\mathbf{x})$. By the single-variable mean value theorem, there exists a $\sigma_{s,t} \in (0, s)$ such that

$$S_{s,t}(\mathbf{x}) = Dg_t(\mathbf{x} + \sigma_{s,t}\mathbf{w})(s\mathbf{w}).$$

But we have

$$Dg_t(\mathbf{x} + \sigma_{s,t}\mathbf{w})(s\mathbf{w}) = Df(\mathbf{x} + \sigma_{s,t}\mathbf{w} + t\mathbf{v})(s\mathbf{w}) - Df(\mathbf{x} + \sigma_{s,t}\mathbf{w})(s\mathbf{w}),$$

so we may apply the mean value theorem to get $\tau_{s,t} \in (0, t)$ such that

$$S_{s,t}(\mathbf{x}) = D^2f(\mathbf{x} + \sigma_{s,t}\mathbf{w} + \tau_{s,t}\mathbf{v})(s\mathbf{w}, t\mathbf{v}).$$

Swapping the roles of $t\mathbf{v}$ and $s\mathbf{w}$ in the previous argument gives $\tau'_{s,t} \in (0, t)$ and $\sigma'_{s,t} \in (0, s)$ such that

$$S_{s,t}(\mathbf{x}) = D^2f(\mathbf{x} + \tau'_{s,t}\mathbf{v} + \sigma'_{s,t}\mathbf{w})(t\mathbf{v}, s\mathbf{w}).$$

Combining these two results and dividing by st gives

$$D^2f(\mathbf{x} + \sigma_{s,t}\mathbf{w} + \tau_{s,t}\mathbf{v})(\mathbf{w}, \mathbf{v}) = D^2f(\mathbf{x} + \tau'_{s,t}\mathbf{v} + \sigma'_{s,t}\mathbf{w})(\mathbf{v}, \mathbf{w}).$$

Since $f \in C^2(U; Y)$, taking the limit as $s, t \rightarrow 0$ gives (6.22).

To prove (6.23), take any $\mathbf{v}_i \in X_i$ and any $\mathbf{w}_j \in X_j$ and apply (6.22) with vectors $\mathbf{v} = (\mathbf{0}, \dots, \mathbf{v}_i, \mathbf{0}, \dots, \mathbf{0})$ and $\mathbf{w} = (\mathbf{0}, \dots, \mathbf{w}_j, \mathbf{0}, \dots, \mathbf{0})$ that are nonzero only in the i th or j th entry, respectively. Finally, (6.24) follows immediately from (6.23). $\square$

Remark 6.6.6. This proposition guarantees that the Hessian matrix of Example 6.6.3 is symmetric.

6.6.2 Higher-Order Directional Derivatives

If $U \subset \mathbb{F}^n$ and $f \in C^k(U; \mathbb{F}^m)$, then $D^k f(\mathbf{x})$ is an element of $\mathcal{B}^k(\mathbb{F}^n; \mathbb{F}^m)$; so it accepts k different vectors from $\mathbb{F}^n$ as inputs, and it returns an element of $\mathbb{F}^m$. It is useful to consider what happens when all of the input vectors are the same vector $\mathbf{v} = \sum_{i=1}^n v_i \mathbf{e}_i$. We can also consider this to be the k th *directional derivative in the direction of $\mathbf{v}$* . We have

$$D_{\mathbf{v}} f(\mathbf{x}) = Df(\mathbf{x})\mathbf{v} = [D_1 f(\mathbf{x}) \quad \dots \quad D_n f(\mathbf{x})]\mathbf{v} = \sum_{j=1}^n v_j D_j f(\mathbf{x}).$$

Repeating the process gives (for convenience we suppress the $\mathbf{x}$)

$$D_{\mathbf{v}}^2 f = D_{\mathbf{v}} D_{\mathbf{v}} f = D_{\mathbf{v}} \sum_{j=1}^n v_j D_j f = \sum_{i=1}^n \sum_{j=1}^n v_i v_j D_i D_j f = \mathbf{v}^T H \mathbf{v},$$

where H is the Hessian of Example 6.6.3.

Iterating k times gives

$$D_{\mathbf{v}}^k f = \sum_{i_1, \dots, i_k=1}^n v_{i_1} \cdots v_{i_k} D_{i_1} \cdots D_{i_k} f. \quad (6.25)$$

We often also write $D_{\mathbf{v}}^k f = D^k f \mathbf{v}^{(k)}$.

Remark 6.6.7. Proposition 6.6.5 shows that (6.25) has repeated terms. If we combine these, we can reexpress (6.25) as

$$D_{\mathbf{v}}^k f = \sum_{j_1+j_2+\dots+j_n=k} \frac{k!}{j_1! j_2! \cdots j_n!} v_1^{j_1} v_2^{j_2} \cdots v_n^{j_n} D_1^{j_1} D_2^{j_2} \cdots D_n^{j_n} f, \quad (6.26)$$

where the sum is taken over all nonnegative integers $j_1, \dots, j_n$ summing to k .

6.6.3 Taylor's Theorem

Recall the very powerful single-variable Taylor's formula.

Theorem 6.6.8 (Taylor's Formula for One Variable). *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a $(k+1)$ -differentiable function. Then, for all $a, h \in \mathbb{R}$, we have*

$$f(a+h) = f(a) + f'(a)h + \frac{f''(a)}{2!}h^2 + \cdots + \frac{f^{(k)}(a)}{k!}h^k + \frac{f^{(k+1)}(c)}{(k+1)!}h^{k+1} \quad (6.27)$$

for some $c \in \mathbb{R}$ between a and $a+h$.

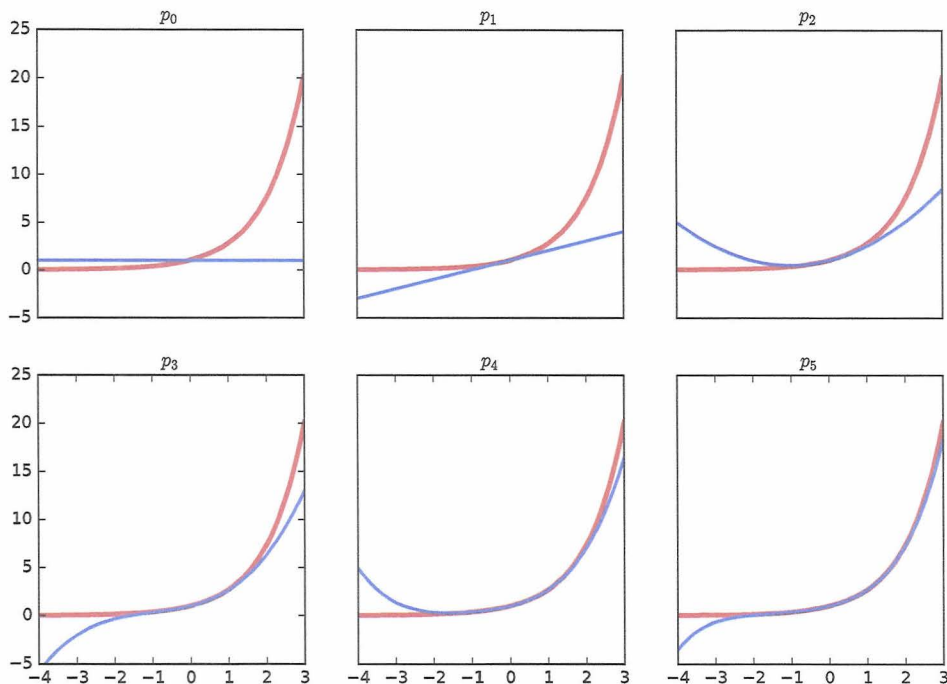


Figure 6.3. Plots of the Taylor polynomials $p_0, \dots, p_5$ for $f(x) = e^x$. The function f is plotted in red, and the Taylor polynomials are plotted in blue. Each polynomial is a good approximation of f in a small neighborhood of $x = 0$, but farther away from 0 the approximation is generally poor. As the degree increases, the neighborhood where the approximation is good increases in size, and the quality of the approximation improves. See Theorem 6.6.8 for the one-dimensional Taylor theorem over $\mathbb{R}$ and Theorem 6.6.9 for the general case.

In this section, we prove the multidimensional version of Taylor's theorem and show how to approximate smooth functions at a point with polynomials.

Theorem 6.6.9. Let $f \in C^k(U; Y)$. If $\mathbf{x} \in U$ and $\mathbf{h} \in X$ are such that the line segment $\ell(\mathbf{x}, \mathbf{x} + \mathbf{h})$ is contained in U , then

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \frac{D^2f(\mathbf{x})\mathbf{h}^{(2)}}{2!} + \cdots + \frac{D^{k-1}f(\mathbf{x})\mathbf{h}^{(k-1)}}{(k-1)!} + R_k, \quad (6.28)$$

where the remainder is given by

$$R_k = \left(\int_0^1 \frac{(1-t)^{k-1}}{(k-1)!} D^k f(\mathbf{x} + t\mathbf{h}) dt \right) \mathbf{h}^{(k)}. \quad (6.29)$$

Proof. We proceed by induction on k . The base case $k = 1$ follows from (6.17). Next, we assume the theorem holds for $k - 1$ and then prove it also holds for k . Thus, assume

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \frac{D^2 f(\mathbf{x})\mathbf{h}^{(2)}}{2!} + \cdots + \frac{D^{k-2} f(\mathbf{x})\mathbf{h}^{(k-2)}}{(k-2)!} + R_{k-1},$$

where

$$R_{k-1} = \left(\int_0^1 \frac{(1-t)^{k-2}}{(k-2)!} D^{k-1} f(\mathbf{x} + t\mathbf{h}) dt \right) \mathbf{h}^{(k-1)}.$$

Note that

$$\begin{aligned} R_{k-1} &= \left(\int_0^1 \frac{(1-t)^{k-2}}{(k-2)!} (D^{k-1} f(\mathbf{x} + t\mathbf{h}) - D^{k-1} f(\mathbf{x})) dt \right) \mathbf{h}^{(k-1)} \\ &\quad + \left(\int_0^1 \frac{(1-t)^{k-2}}{(k-2)!} D^{k-1} f(\mathbf{x}) dt \right) \mathbf{h}^{(k-1)} \\ &= \left(\int_0^1 \frac{(1-t)^{k-2}}{(k-2)!} \left(\int_0^t D(D^{k-1} f(\mathbf{x} + u\mathbf{h}))\mathbf{h} du \right) dt \right) \mathbf{h}^{(k-1)} \\ &\quad + \left(\frac{D^{k-1} f(\mathbf{x})\mathbf{h}^{(k-1)}}{(k-2)!} \right) \int_0^1 (1-t)^{k-2} dt, \end{aligned}$$

where the first term of the last equality follows from the integral mean value theorem (6.17) and the change of variables formula (6.19).

Simplifying and changing the order of integration gives

$$\begin{aligned} R_{k-1} &= \frac{D^{k-1} f(\mathbf{x})\mathbf{h}^{(k-1)}}{(k-1)!} + \left(\int_0^1 D^k f(\mathbf{x} + u\mathbf{h}) \left(\int_u^1 \frac{(1-t)^{k-2}}{(k-2)!} dt \right) du \right) \mathbf{h}^{(k)} \\ &= \frac{D^{k-1} f(\mathbf{x})\mathbf{h}^{(k-1)}}{(k-1)!} + \left(\int_0^1 \frac{(1-u)^{k-1}}{(k-1)!} D^k f(\mathbf{x} + u\mathbf{h}) du \right) \mathbf{h}^{(k)}. \end{aligned}$$

Thus, (6.28) holds, and the proof is complete. $\square$

Remark 6.6.10. For each $k \geq 0$, neglecting the remainder term gives the degree- k Taylor polynomial approximation of f . If f is smooth and the remainder R_k can be shown to go to zero as $k \rightarrow \infty$, then the Taylor series converges to f .

Example 6.6.11. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = e^{x+y}$. We find the second-order Taylor polynomial, that is, the Taylor polynomial of degree 2 of f at $(0, 0)$. We have

$$D_1 f(x, y) = D_2 f(x, y) = e^{x+y}.$$

Likewise all second-order partials are equal to e^{x+y} , and

$$D_1 D_1 f(x, y) = D_2 D_1 f(x, y) = D_1 D_2 f(x, y) = D_2 D_2 f(x, y) = e^{x+y}.$$

All of these derivative terms evaluate to 1 at $\mathbf{0}$. Thus, the second-order Taylor approximation of f at $\mathbf{0}$, evaluated at $\mathbf{x} = (x, y)$, is

$$T_2(\mathbf{x}) = \sum_{j=0}^2 \frac{D_{\mathbf{x}}^j f(\mathbf{0})}{j!} = f(\mathbf{0}) + D_{\mathbf{x}}f(\mathbf{0}) + \frac{D_{\mathbf{x}}^2 f(\mathbf{0})}{2},$$

where $f(\mathbf{0}) = 1$,

$$D_{\mathbf{x}}f(\mathbf{0}) = xD_1f(0,0) + yD_2f(0,0) = x + y \text{ and}$$

$$D_{\mathbf{x}}^2 f(\mathbf{0}) = x^2 D_1^2 f(0,0) + 2xy D_1 D_2 f(0,0) + y^2 D_2^2 f(0,0) = x^2 + 2xy + y^2.$$

It follows that

$$T_2(x, y) = 1 + x + y + \frac{1}{2}(x^2 + 2xy + y^2).$$

Example 6.6.12. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = \cos(x)e^{3y}$. The first- and second-order partial derivatives at $(0, 0)$ are

$$\begin{aligned} f(0, 0) &= 1, & D_{11}f(0, 0) &= -\cos(x)e^{3y}|_{(0,0)} = -1, \\ D_1f(0, 0) &= -\sin(x)e^{3y}|_{(0,0)} = 0, & D_{12}f(0, 0) &= -3\sin(x)e^{3y}|_{(0,0)} = 0, \\ D_2f(0, 0) &= 3\cos(x)e^{3y}|_{(0,0)} = 3, & D_{21}f(0, 0) &= -3\sin(x)e^{3y}|_{(0,0)} = 0, \\ & & D_{22}f(0, 0) &= 9\cos(x)e^{3y}|_{(0,0)} = 9. \end{aligned}$$

Thus, the second-order Taylor polynomial is

$$\begin{aligned} T_2(x, y) &= f(0, 0) + [D_1f(0, 0) \quad D_2f(0, 0)] \begin{bmatrix} x \\ y \end{bmatrix} \\ &\quad + \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} D_{11}f(0, 0) & D_{12}f(0, 0) \\ D_{21}f(0, 0) & D_{22}f(0, 0) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= 1 + [0 \quad 3] \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= 1 + 3y - \frac{1}{2}x^2 + \frac{9}{2}y^2. \end{aligned}$$

Example 6.6.13. For general functions $z : \mathbb{R}^2 \rightarrow \mathbb{R}$, the first-order Taylor polynomial approximation at $\mathbf{x}_0 = (x_0, y_0)$ is the plane tangent to the graph of z at $(x_0, y_0, z(x_0, y_0))$:

$$z(x, y) = z(x_0, y_0) + \frac{\partial z}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial z}{\partial y}(x_0, y_0)(y - y_0). \quad (6.30)$$

As a corollary to Taylor's theorem, we have the following,

Corollary 6.6.14. *If $\|D^k f(\mathbf{x} + t\mathbf{h})\| < M$ for all $t \in [0, 1]$, then the remainder R_k is bounded by*

$$\|R_k\|_Y \leq \frac{M}{k!} \|\mathbf{h}\|_X^k. \quad (6.31)$$

Proof. The proof is Exercise 6.33. $\square$

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

6.1. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{x^3 y}{x^6 + y^2}, & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0). \end{cases}$$

Show that the partial derivatives of f exist at $(0, 0)$ but are discontinuous there. Hint: To show the discontinuity, consider a sequence of the form $(a/n, b/n^3)$.

6.2. Prove Proposition 6.1.9.

6.3. Let $A \subset \mathbb{R}^2$ be an open, path-connected set, and let $f : A \rightarrow \mathbb{R}$. If for each $a \in A$ we have $D_1 f(a) = D_2 f(a) = 0$, prove that f is constant on A . Hint: Consider first the case where $A = B(\mathbf{x}, r)$ is an open ball.

6.4. Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by $f(x, y) = xye^{x+y}$.

(i) Find the directional derivative of f in the direction $\mathbf{u} = (\frac{3}{5}, \frac{4}{5})$ at the point $(1, -1)$.

(ii) Find the direction in which f is increasing the fastest at $(1, -1)$.

6.5.† Let $A \subset \mathbb{R}^2$ be an open set and $f : A \rightarrow \mathbb{R}$. If the partial derivatives of f exist and are bounded on A , then prove that f is continuous on A .

6.6. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{xy^2}{x^2 + y^2}, & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0). \end{cases}$$

Show that f is continuous but not differentiable at $(0, 0)$.

6.7. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2}, & x^2 + y^2 \neq 0, \\ 0, & x^2 + y^2 = 0. \end{cases}$$

Show that the partial derivatives exist at $(0, 0)$ but that f is discontinuous at $(0, 0)$, so f cannot be differentiable there.

6.8. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{x|y|}{\sqrt{x^2 + y^2}}, & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0). \end{cases}$$

Show that f has a directional derivative in every direction at $(0, 0)$ but is not differentiable there.

6.9. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} (x^2 + y^2) \sin\left(\frac{1}{\sqrt{x^2 + y^2}}\right), & (x, y) \neq (0, 0), \\ 0, & (x, y) = (0, 0). \end{cases}$$

Show that f is differentiable at $(0, 0)$. Then show that the partial derivatives are bounded near $(0, 0)$ but are discontinuous there.

6.10. Use the definition of the derivative to prove the claims made in Remark 6.2.6:

- (i) Using row vectors to represent $(\mathbb{R}^m)^*$, if $A \in M_{n \times m}(\mathbb{R})$ defines a function $g : \mathbb{R}^n \rightarrow (\mathbb{R}^m)^*$ given by $g(\mathbf{x}) = \mathbf{x}^\top A$, then $Dg : \mathbb{R}^n \rightarrow (\mathbb{R}^m)^*$ is the linear transformation $\mathbf{v} \mapsto \mathbf{v}^\top A$.
- (ii) Using column vectors to represent $(\mathbb{R}^m)^*$, prove that the function g is given by $g(\mathbf{x}) = A^\top \mathbf{x}$ and that Dg is the linear transformation $\mathbf{v} \mapsto A^\top \mathbf{v}$.

6.11. Let $X = C([0, 1]; \mathbb{R})$ with the sup norm.

- (i) Fix $p \in [0, 1]$ and let $e_p : X \rightarrow \mathbb{R}$ be given by $e_p(f) = f(p)$. Prove that e_p is a linear transformation. Use the definition of the derivative to verify that $De_p(f)(g) = e_p(g)$ for any $f, g \in X$.
- (ii) Let $K : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$ be a continuous function, and let $\mathcal{K} : X \rightarrow X$ be given by $\mathcal{K}[f](s) = \int_0^1 K(s, t)f(t) dt$. Verify that $\mathcal{K}$ is a linear transformation. Use the definition of derivative to verify that $D\mathcal{K}(f)(g) = \mathcal{K}[g]$.

These are both special cases of Example 6.3.3, which says that for any linear transformation $L : U \rightarrow Y$, the derivative $DL(\mathbf{x})$ is equal to L for every $\mathbf{x}$.

- 6.12. Let $X = C([0, 1]; \mathbb{R})$ with the sup norm, and let the function $T : X \rightarrow \mathbb{R}$ be given by $T(f) = \int_0^1 t^2 f^2(t) dt$. Let $L : X \rightarrow \mathbb{R}$ be the function $L(g) = \int_0^1 2t^2 f(t)g(t) dt$. Prove that L is a linear transformation. Use the definition of the Fréchet derivative to prove that $DT(f)(g) = L(g)$ for all $f \in X$.
- 6.13. Let X be a Banach space and $A \in \mathcal{B}(X)$ be a bounded linear operator on X . Let $At : \mathbb{R} \rightarrow \mathcal{B}(X)$ be scalar multiplication of A by t , and let $E : \mathbb{R} \rightarrow \mathcal{B}(X)$ be the exponential of At , given by $E(t) = e^{At}$ (see Example 5.7.2). Use the definition of the Fréchet derivative to prove that

$$DE(t)(s) = Ae^{At}s$$

for every $t, s \in \mathbb{R}$. You may assume without proof that $E(t+s) = E(t)E(s)$.

- 6.14. Prove Corollary 6.3.8.
- 6.15. Prove Proposition 6.3.11.

- 6.16.  Prove Proposition 6.4.6.
- 6.17. Let $\gamma(t)$ be a differentiable curve in $\mathbb{R}^n$. If there is some differentiable function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ with $F(\gamma(t)) = C$ constant, show that $DF(\gamma(t))^T$ is orthogonal to the tangent vector $\gamma'(t)$.
- 6.18. Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ and $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$f(x, y) = (\sin(y) - x, e^x - y) \quad \text{and} \quad g(x, y) = (xy, x^2 + y^2).$$

Compute $D(g \circ f)(0, 0)$.

- 6.19. In the previous problem, compute $D(\langle f, g \rangle)(0, 0)$, where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^2$.
- 6.20. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be defined by

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2,$$

where A is an $n \times n$ matrix and $\mathbf{b} \in \mathbb{R}^n$. Find $Df(\mathbf{x}_0)$.

- 6.21.*† Let $\mathcal{L} : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a bounded integrable function with continuous partial derivative $D_2\mathcal{L}$ satisfying the additional condition that for every $\varepsilon > 0$ and for every compact subset $K \subset \mathbb{R}^2$, there is a (uniform) $\delta > 0$ such that if $|h| < \delta$, then for every $(t, s) \in K$ we have

$$|\mathcal{L}(t, s+h) - \mathcal{L}(t, s) - D_2\mathcal{L}(t, s)h| \leq \varepsilon|h|.$$

Consider the functional $S : L^\infty([a, b]; \mathbb{R}) \rightarrow \mathbb{R}$ defined by

$$S(u) = \int_a^b \mathcal{L}(t, u(t)) dt.$$

Prove that the Fréchet derivative $DS(u)$ is given by

$$DS(u)(v) = \int_a^b D_2\mathcal{L}(t, u(t))v dt \quad (6.32)$$

using the following steps:

- (i) Prove that for every $u \in C([a, b]; \mathbb{R})$ and for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every $t \in [a, b]$, if $|h(t)| < \delta$, then we have

$$|\mathcal{L}(t, u(t) + h(t)) - \mathcal{L}(t, u(t)) - D_2\mathcal{L}(t, u(t))h(t)| \leq \varepsilon \|h\|_\infty.$$

- (ii) Use properties of the integral (Proposition 5.10.12) to prove that if $\|\xi\|_{L^\infty} < \delta$, then

$$|S(u + \xi) - S(u) - \int_a^b D_2\mathcal{L}(t, u(t))\xi dt| \leq \varepsilon \|\xi\|(b - a).$$

Conclude by Lemma 6.4.1 that the Fréchet derivative of S is given by (6.32).

6.22. Prove Corollary 6.5.5. Hint: Consider the function $g(t) = f(t\mathbf{y} + (1 - t)\mathbf{x})$.

6.23. Prove Corollary 6.5.6. Hint: Consider the function $F(t) = \int_{g(c)}^t f(\tau) d\tau$.

- (i) First prove the result assuming the chain rule holds even at points where $g(s) = c$ or $g(s) = d$.

- (ii)* Prove that the chain rule holds for the extension of $D(\text{Fog})$ to points where $g(s) = c$ or $g(s) = d$.

6.24. Prove the claims in Remark 6.5.8: Given an open subset U of a finite-dimensional Banach space and a sequence $(f_n)_{n=0}^\infty$ in $C(U; Y)$, prove the following:

- (i) The sequence $(f_n)_{n=0}^\infty$ is uniformly convergent on compact subsets in U if and only if the restriction of $(f_n)_{n=0}^\infty$ to the ball $\overline{B(\mathbf{x}, r)}$ is uniformly convergent for every closed ball $\overline{B(\mathbf{x}, r)}$ in U .

- (ii) If $U = B(\mathbf{x}_0, R)$, then prove that $(f_n)_{n=0}^\infty$ is uniformly convergent on compact subsets if and only if for every $0 < r < R$ the restriction of $(f_n)_{n=0}^\infty$ to the closed ball $\overline{B(\mathbf{x}_0, r)}$ centered at $\mathbf{x}_0$ converges uniformly. Hint: For any compact subset K , show that $d(K, U^c) > 0$, and use this fact to construct a closed ball $\overline{B(\mathbf{x}_0, r)} \subset U$ containing K .

- (iii)* The sequence $(f_n)_{n=0}^\infty$ is Cauchy on U (see Definition 6.5.7) if and only if the restriction of $(f_n)_{n=0}^\infty$ to the ball $\overline{B(\mathbf{x}, r)}$ is Cauchy for every closed ball $\overline{B(\mathbf{x}, r)}$ in U .

- (iv)* If $U = B(\mathbf{x}_0, R)$, then prove that $(f_n)_{n=0}^\infty$ is Cauchy on U if and only if for every $0 < r < R$ the restriction of $(f_n)_{n=0}^\infty$ to $\overline{B(\mathbf{x}_0, r)}$ is Cauchy.

6.25. For each integer $n \geq 1$, let $f_n : [-1, 1] \rightarrow \mathbb{R}$ be $f_n(x) = \sqrt{\frac{1}{n^2} + x^2}$.

- (i) Prove that each f_n is differentiable on $(-1, 1)$.

- (ii) Prove that $(f_n)_{n=0}^\infty$ converges uniformly to $f(x) = |x|$.

- (iii) Prove that f is not differentiable at 0.

- (iv) Explain why this does not contradict Theorem 6.5.11.

6.26. For any $a > 0$, for x in the interval (a, ∞) , and for any $n \in \mathbb{N}$, show that

$$\lim_{N \rightarrow \infty} \frac{d^n}{dx^n} \left(\frac{1 - e^{-Nx}}{x} \right) = (-1)^n \frac{n!}{x^{n+1}}.$$

6.27.  Let $U \subset X$ be the open ball $B(\mathbf{x}_0, r)$. Assume that $(f_n)_{n=1}^\infty$ is a sequence in $(C^1(U; Y), \|\cdot\|_{L^\infty})$ such that the series $\sum_{n=0}^\infty Df_n$ converges absolutely (using the sup norm) on all compact subsets of U . Assume also that $\sum_{n=0}^\infty f_n(\mathbf{x}_0)$ converges (as a series in Y). Prove that the sum $\sum_{n=0}^\infty f_n$ converges uniformly on compact subsets in U and that the derivative commutes with the sum:

$$D \sum_{n=0}^\infty f_n = \sum_{n=0}^\infty Df_n.$$

6.28. Let $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ be twice continuously differentiable. If $x = r \cos \theta$ and $y = r \sin \theta$, show that the Laplacian $\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ takes the form

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}.$$

- 6.29.  Prove that if X is a normed linear space, if $U \subset X$ is an open subset, and if $f : U \rightarrow \mathbb{R}$ is a differentiable function that attains a local minimum or a local maximum at $\mathbf{x} \in U$, then $Df(\mathbf{x}) = 0$. This is explored much more in Volume 2.
- 6.30. Find the second-order Taylor polynomial for $f(x, y) = \cos(xy)$ at the point $(0, 0)$.
- 6.31. Find the second-order Taylor polynomial for $g(x, y, z) = e^{2x+yz}$ at the point $(0, 0, 0)$.
- 6.32. Find the second-order Taylor polynomial for $f(x, y) = \log(1 + x \sin(y))$ at $(0, 0)$. Now compute the second-order Taylor polynomial of $g(t) = \log(1 + t)$ at $t = 0$ and the first-order Taylor polynomial of $r(y) = \sin(y)$ at $y = 0$ and combine them to get a polynomial for $f(x, y) = g(xr(y))$. Compare the results.
- 6.33. Prove Corollary 6.6.14.
- 6.34.  Assume $U \subset \mathbb{F}^n$ is an open subset and $f \in C^2(U; \mathbb{R})$. Prove that if $\mathbf{x}_0 \in U$ is such that $Df(\mathbf{x}_0) = 0$ and $D^2 f(\mathbf{x}_0)$ is positive definite (see Definition 4.5.1), then f attains a local minimum at $\mathbf{x}_0$. You may assume that there is a neighborhood $B(\mathbf{x}_0, \delta)$ of $\mathbf{x}_0$ where every eigenvalue is greater than ε , and hence $\mathbf{v}^T D^2 f(\mathbf{x}) \mathbf{v} > \varepsilon \|\mathbf{v}\|$ for all $\mathbf{v}$ and for all $\mathbf{x}$ in $B(\mathbf{x}_0, \delta)$ (this follows from the implicit function theorem, which is proved in the next chapter). Hint: Consider using Taylor's theorem.

7

Contraction Mappings and Applications

There are fixed points through time where things must always stay the way they are. This is not one of them. This is an opportunity.

—Dr. Who

Fixed-point theorems are among the most powerful tools in mathematical analysis. They are found in nearly every area of pure and applied mathematics. In Corollary 5.9.16, we saw a very simple example of a fixed-point theorem, namely, if a map $f : [a, b] \rightarrow [a, b]$ is continuous, then it has a fixed point. This result generalizes to higher dimensions in what is called the Brouwer fixed-point theorem, which states that a continuous map on the closed unit ball $D(0, 1) \subset \mathbb{F}^n$ into itself has a fixed point. The Brouwer fixed-point theorem can be generalized further to infinite dimensions by the Leray–Schauder fixed-point theorem, which is important in both functional analysis and partial differential equations.

Most fixed-point theorems only say when a fixed point exists, and do not give any additional information on how to find a fixed point or how many there are. In this chapter, we study the contraction mapping principle, which gives conditions guaranteeing both the existence and uniqueness of a fixed point. Moreover, it also provides a way to actually compute the fixed point using the *method of successive approximations*. As a result, the contraction mapping principle is widely used in pure and applied mathematics and is the basis of many computational algorithms. Chief among these are the Newton family of algorithms, which includes Newton’s method for finding zeros of a function, and several close cousins called *quasi-Newton methods*. These numerical methods are ubiquitous in applications, particularly in *optimization problems* and *inverse problems*.

The contraction mapping principle also gives two very important theorems: the implicit function theorem and the inverse function theorem. Given a system of equations, these two theorems give straightforward criteria that guarantee the existence of a function (an implicit function) solving the system. They also allow us to differentiate these implicit functions without ever explicitly writing the functions down. These two theorems are essential tools in differential geometry, optimization, and differential equations, as well as in applications like economics and physics.

7.1 Contraction Mapping Principle

In this section, we prove the contraction mapping principle and provide several examples of its use in applications.

Definition 7.1.1. A point $\bar{x} \in X$ is a fixed point of the function $f : X \rightarrow X$ if $f(\bar{x}) = \bar{x}$.

Example 7.1.2. Let $f : [0, 1] \rightarrow [0, 1]$ be defined by $f(x) = x^n$ for some positive n ; then 0 and 1 are both fixed points of f .

Our first theorem tells us that a class of functions called *contraction mappings* always have a unique fixed point.

Definition 7.1.3. Assume D is a subset of a normed linear space $(X, \|\cdot\|)$. The function $f : D \rightarrow D$ is a contraction mapping if there exists $0 \leq k < 1$ such that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq k\|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in D. \quad (7.1)$$

Remark 7.1.4. It is easy to see that contraction mappings are continuous. In fact, they are Lipschitz continuous with constant k (see Example 5.2.2(iii)).

Example 7.1.5. Consider the mapping $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = \frac{x}{2} + 10$. For any $x, y \in \mathbb{R}$ we have

$$|f(x) - f(y)| = \left| \frac{x}{2} + 10 - \frac{y}{2} - 10 \right| = \frac{1}{2}|x - y|.$$

Thus, $f(x)$ is a contraction mapping with contraction factor $k = \frac{1}{2}$ and fixed point $x = 20$.

Example 7.1.6. Let μ be a positive real number, and let $T : [0, 1] \rightarrow [0, 1]$ be defined by $T(x) = 4\mu x(1 - x)$. This function is important in population dynamics. For any $x, y \in [0, 1]$, we have $-1 \leq (1 - x - y) \leq 1$, and therefore

$$\begin{aligned} |T(x) - T(y)| &= 4\mu|x(1 - x) - y(1 - y)| = 4\mu|x - y - (x^2 - y^2)| \\ &= 4\mu|x - y||1 - x - y| \leq 4\mu|x - y|. \end{aligned}$$

If $4\mu < 1$, then T is a contraction mapping.

Unexample 7.1.7. Let $f : (0, 1) \rightarrow (0, 1)$ be given by $f(x) = x^2$. For any x we have $|f(x) - f(0)| = |x^2| < |x - 0|$; but f is not a contraction mapping on $(0, 1)$ because for any distinct $x, y \in (1/2, 1)$ we have $|f(x) - f(y)| = |x^2 - y^2| = |x - y||x + y| > |x - y|$.

Theorem 7.1.8 (Contraction Mapping Principle). Assume D is a nonempty closed subset of a Banach space $(X, \|\cdot\|)$. If $f : D \rightarrow D$ is a contraction mapping, then there exists a unique fixed point $\bar{x} \in D$ of f .

Proof. Let $\mathbf{x}_0 \in D$. Iteratively define the sequence $(\mathbf{x}_n)_{n=0}^\infty$ by the rule

$$\mathbf{x}_{n+1} = f(\mathbf{x}_n), \quad n \in \mathbb{N}. \quad (7.2)$$

We first prove that the sequence is Cauchy. Since f is a contraction on D , say with constant k , it follows that

$$\begin{aligned} \|\mathbf{x}_n - \mathbf{x}_{n-1}\| &= \|f(\mathbf{x}_{n-1}) - f(\mathbf{x}_{n-2})\|, \\ &\leq k\|\mathbf{x}_{n-1} - \mathbf{x}_{n-2}\|, \\ &\vdots \\ &\leq k^{n-1}\|\mathbf{x}_1 - \mathbf{x}_0\|. \end{aligned}$$

Hence,

$$\begin{aligned} \|\mathbf{x}_n - \mathbf{x}_m\| &\leq \|\mathbf{x}_n - \mathbf{x}_{n-1}\| + \cdots + \|\mathbf{x}_{m+1} - \mathbf{x}_m\| \\ &\leq k^{n-1}\|\mathbf{x}_1 - \mathbf{x}_0\| + \cdots + k^m\|\mathbf{x}_1 - \mathbf{x}_0\| \\ &= k^m(1 + k + \cdots + k^{n-m-1})\|\mathbf{x}_1 - \mathbf{x}_0\| \\ &\leq \frac{k^m}{1-k}\|\mathbf{x}_1 - \mathbf{x}_0\|. \end{aligned} \quad (7.3)$$

Given $\varepsilon > 0$, we can choose N such that $\frac{k^N}{1-k}\|\mathbf{x}_1 - \mathbf{x}_0\| < \varepsilon$. It follows that $\|\mathbf{x}_n - \mathbf{x}_m\| < \varepsilon$ whenever $n > m \geq N$. Therefore, the sequence $(\mathbf{x}_n)_{n=0}^\infty$ is Cauchy.

Since X is complete and D is closed, the sequence converges to some $\bar{x} \in D$. To prove that $f(\bar{x}) = \bar{x}$, let $\varepsilon > 0$ and choose $N > 0$ so that $\|\bar{x} - \mathbf{x}_n\| < \varepsilon/2$ whenever $n \geq N$. Thus,

$$\begin{aligned} \|f(\bar{x}) - \bar{x}\| &\leq \|f(\bar{x}) - f(\mathbf{x}_n)\| + \|\bar{x} - f(\mathbf{x}_n)\| \\ &\leq k\|\bar{x} - \mathbf{x}_n\| + \|\bar{x} - \mathbf{x}_{n+1}\| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Since ε is arbitrary, we have $f(\bar{x}) = \bar{x}$.

To show uniqueness, suppose $\bar{y} \in D$ is some other fixed point of f . Then

$$\|\bar{x} - \bar{y}\| = \|f(\bar{x}) - f(\bar{y})\| \leq k\|\bar{x} - \bar{y}\| < \|\bar{x} - \bar{y}\|,$$

which is a contradiction. $\square$

Remark 7.1.9. The contraction mapping principle can be proved for complete metric spaces instead of just Banach spaces without any extra work. Simply change each occurrence of $\|\mathbf{x} - \mathbf{y}\|$ to $d(\mathbf{x}, \mathbf{y})$ in the proof above.

Remark 7.1.10. The proof of the contraction mapping principle above gives an algorithm for finding the unique fixed point, given by (7.2). This is called the *method of successive approximations*. To use it, pick *any* initial guess $\mathbf{x}_0 \in D$, and the sequence $f(\mathbf{x}_0), f^2(\mathbf{x}_0), \dots$ converges to the fixed point. Taking the limit of (7.3) as $n \rightarrow \infty$ shows that the error of the m th approximation $f^m(\mathbf{x}_0)$ is at most $\frac{k^m}{1-k} \|\mathbf{x}_1 - \mathbf{x}_0\|$.

Vista 7.1.11. The contraction mapping principle plays a key role in Blackwell's theorem, which is fundamental to solving sequential decision-making problems (dynamic optimization). We discuss this application in Volume 2. The contraction mapping principle is also important in the proof of existence and uniqueness of solutions to ordinary differential equations. This is covered in Volume 4.

Example 7.1.12. We can use the method of successive approximations to compute positive square roots of numbers greater than 1. If $c^2 = b > 1$, then c is a fixed point of the mapping $f(x) = \frac{1}{2}(x + \frac{b}{x})$. The function f is not a contraction mapping on all of $\mathbb{R}$, but the interval $[\sqrt{b/2}, \infty)$ is closed in $\mathbb{R}$ and f maps $[\sqrt{b/2}, \infty)$ to itself. To see that f is a contraction mapping on $[\sqrt{b/2}, \infty)$, we compute

$$\begin{aligned} |f(x) - f(y)| &= \frac{1}{2} \left| x - y + \frac{b}{x} - \frac{b}{y} \right| \\ &\leq \frac{1}{2} |x - y| \left| 1 - \frac{b}{xy} \right|. \end{aligned}$$

So when the domain of f is restricted to $[\sqrt{b/2}, \infty)$ we have

$$|f(x) - f(y)| \leq \frac{1}{2} |x - y| \left| 1 - \frac{b}{b/2} \right| = \frac{1}{2} |x - y|.$$

Therefore f is a contraction mapping on this domain with $\sqrt{b}$ as its fixed point. Starting with any point in $[\sqrt{b/2}, \infty)$, the method of successive approximations converges (fairly rapidly) to $\sqrt{b}$.

Example 7.1.13. Recall that $(C([a, b]; \mathbb{F}), \|\cdot\|_{L^\infty})$ is a Banach space. Consider the operator $L : C([a, b]; \mathbb{F}) \rightarrow C([a, b]; \mathbb{F})$, given by

$$L[f](x) = \lambda \int_a^b K(x, y) f(y) dy$$

on $C([a, b]; \mathbb{F})$, where K is continuous and satisfies $|K(x, y)| \leq M$ on the domain $[a, b] \times [a, b]$.

This operator L is called a *Fredholm integral transform with kernel*^a K , and it arises in many applications. For example, if $K(x, y) = \frac{e^{-ixy}}{2\pi}$ and $a = 0$ and $b = 2\pi$, then L is the Fourier transform.

The map L is a contraction for sufficiently small $|\lambda|$. More specifically, we have

$$\begin{aligned} \|L[f] - L[g]\|_{L^\infty} &= \|L[f - g]\|_{L^\infty} \\ &= \left\| \lambda \int_a^b K(x, y) (f(y) - g(y)) dy \right\|_{L^\infty} \\ &\leq |\lambda| M(b - a) \|f - g\|_{L^\infty}. \end{aligned}$$

Thus, if $|\lambda| < \frac{1}{M(b-a)}$, then L is a contraction on $C([a, b]; \mathbb{F})$, and there exists a unique function $f(x) \in C([a, b]; \mathbb{R})$ such that

$$f(x) = \lambda \int_a^b K(x, y) f(y) dy.$$

^aThe word *kernel* here has nothing to do with kernels of linear transformations nor with kernels of group homomorphisms. It is unfortunate that standard mathematical usage has given the same word two entirely different meanings, but the meaning is usually clear from the context.

It sometimes happens that a function f is not a contraction mapping, but an iterate f^n is a contraction mapping. In this case, f still has a unique fixed point.

Theorem 7.1.14. Assume D is a nonempty closed subset of the Banach space $(X, \|\cdot\|)$. If $f : D \rightarrow D$ and f^n is a contraction mapping for some positive integer n , then there exists a unique fixed point $\bar{x} \in D$ of f .

Proof. The proof is Exercise 7.6. $\square$

7.2 Uniform Contraction Mapping Principle

We now generalize the results of the previous section to the *uniform* contraction mapping principle. This is one of the more challenging theorems in the text, but it is also very powerful. In Section 7.4, we show that the implicit and inverse function theorems are just corollaries of the uniform contraction mapping principle.

Definition 7.2.1. If D is a nonempty subset of a normed linear space $(X, \|\cdot\|_X)$ and B is some arbitrary set, then the function $f : D \times B \rightarrow D$ is called a uniform contraction mapping if there exists $0 \leq \lambda < 1$ such that

$$\|f(\mathbf{x}_2, b) - f(\mathbf{x}_1, b)\|_X \leq \lambda \|\mathbf{x}_2 - \mathbf{x}_1\|_X \quad (7.4)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and all $b \in B$.

Example 7.2.2. Example 7.1.5 can be generalized to a uniform contraction mapping. Let $g : \mathbb{R} \times \mathbb{N} \rightarrow \mathbb{R}$ be given by $g(x, b) = \frac{x}{2} + b$. For x_1 and x_2 in $\mathbb{R}$ we have

$$\|g(x_2, b) - g(x_1, b)\| = \left\| \frac{x_2}{2} + b - \left(\frac{x_1}{2} + b \right) \right\| = \frac{1}{2} \|x_2 - x_1\|.$$

Hence, g is a uniform contraction mapping with $\lambda = \frac{1}{2}$. Although λ is independent of b , the fixed point is not. For example, when $b = 10$ the fixed point is 20, but when $b = 20$ the fixed point is 40.

Example 7.2.3. Let $B = D = [1, 2]$, and let $f : D \times B \rightarrow D$ be given by $f(x, b) = \frac{1}{2}(x + \frac{b}{x})$. The reader should check that $f(x, b) \in D$ for all $(x, b) \in D \times B$. Using arguments similar to those in Example 7.1.12, we compute for any $b \in B$ and any $x_1, x_2 \in D$ that

$$\begin{aligned} |f(x_1, b) - f(x_2, b)| &= \frac{1}{2} |x_1 - x_2| \left| 1 - \frac{b}{x_1 x_2} \right| \\ &\leq \frac{1}{2} |x_1 - x_2|. \end{aligned}$$

Therefore, f is a uniform contraction mapping on $D \times B$.

7.2.1 The Uniform Contraction Mapping Principle

A uniform contraction $f : D \times B \rightarrow D$ can be thought of as a family of contraction mappings—one for each $b \in B$. Each one of these contractions has a unique fixed point, so this gives a map $g : B \rightarrow D$, sending b to the unique fixed point of the corresponding contraction. The uniform contraction mapping principle gives conditions that guarantee the function g is differentiable. This is useful because it gives us a way to construct new differentiable functions with various desirable properties.

Theorem 7.2.4 (Uniform Contraction Mapping Principle). Assume that $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces, $U \subset X$ and $V \subset Y$ are open, and the function $f : \overline{U} \times V \rightarrow \overline{U}$ is a uniform contraction with constant $0 \leq \lambda < 1$. Define a function $g : V \rightarrow \overline{U}$ that sends each $\mathbf{y} \in V$ to the unique fixed point of the contraction $f(\cdot, \mathbf{y})$. If $f \in C^k(\overline{U} \times V; \overline{U})$ for some $k \in \mathbb{N}$, then $g \in C^k(V; \overline{U})$.

Remark 7.2.5. Although we have only defined derivatives on open subsets of Banach spaces, the theorem requires f to be C^k on the set $\overline{U} \times V$, which is not open. For our purposes here, derivatives and differentiability at a point $\mathbf{x}$ that is not interior to the domain mean that the limit (6.7) defining the derivative is only taken with respect to those $\mathbf{h}$ with $\mathbf{x} + \mathbf{h}$ in the domain.

In order to prove the uniform contraction mapping principle we first establish several lemmata. The rough idea is to first prove continuity and differentiability of g , and then induct on k . Continuity of g and its derivative (if it exists) is the easy part of the proof, and the existence of the derivative is much harder. We begin by observing that the definition of g gives

$$g(\mathbf{y}) = f(g(\mathbf{y}), \mathbf{y}) \quad (7.5)$$

for every $\mathbf{y} \in V$.

Lemma 7.2.6. *Let X, Y, U, V, f , and g be as in the hypothesis of Theorem 7.2.4 for $k = 0$. If f is continuous at each $(g(\mathbf{y}), \mathbf{y}) \in \overline{U} \times V$, then g is continuous at each $\mathbf{y} \in V$.*

Proof. Given $\varepsilon > 0$, choose $B(\mathbf{y}, \delta) \subset V$ so that

$$\|f(g(\mathbf{y}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y})\|_X < (1 - \lambda)\varepsilon$$

whenever $\|\mathbf{k}\|_Y < \delta$. Thus,

$$\begin{aligned} \|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y})\|_X &= \|f(g(\mathbf{y} + \mathbf{k}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y})\|_X \\ &\leq \|f(g(\mathbf{y} + \mathbf{k}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y} + \mathbf{k})\|_X + \|f(g(\mathbf{y}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y})\|_X \\ &< \lambda\|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y})\|_X + (1 - \lambda)\varepsilon, \end{aligned}$$

which implies that $\|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y})\|_X < \varepsilon$ whenever $\|\mathbf{k}\|_Y < \delta$. Therefore, g is continuous at $\mathbf{y}$. $\square$

Lemma 7.2.7. *Let X, Y, U, V, f , and g be as in the hypothesis of Theorem 7.2.4 for $k = 1$. If f is C^1 at each $(g(\mathbf{y}), \mathbf{y}) \in \overline{U} \times V$, then the function $\phi : \mathcal{B}(Y; X) \times V \rightarrow \mathcal{B}(Y; X)$ defined by*

$$\phi(A, \mathbf{y}) = D_1 f(g(\mathbf{y}), \mathbf{y})A + D_2 f(g(\mathbf{y}), \mathbf{y}) \quad (7.6)$$

is a uniform contraction.

Proof. By Proposition 6.3.10, we have that $\|D_1 f(g(\mathbf{y}), \mathbf{y})\| \leq \lambda$ for each $\mathbf{y} \in V$, where $\|\cdot\|$ is the induced norm,⁴¹ and so

$$\begin{aligned} \|\phi(A_2, \mathbf{y}) - \phi(A_1, \mathbf{y})\| &= \|D_1 f(g(\mathbf{y}), \mathbf{y})A_2 + D_2 f(g(\mathbf{y}), \mathbf{y}) - D_1 f(g(\mathbf{y}), \mathbf{y})A_1 - D_2 f(g(\mathbf{y}), \mathbf{y})\| \\ &\leq \|D_1 f(g(\mathbf{y}), \mathbf{y})\| \|A_2 - A_1\| \\ &\leq \lambda \|A_2 - A_1\|. \quad \square \end{aligned}$$

⁴¹Note that the conclusion of Proposition 6.3.10 still holds for $f \in C^1(\overline{U} \times V)$.

Remark 7.2.8. The previous lemma shows that for each $\mathbf{y} \in V$, there exists a unique fixed point $Z(\mathbf{y}) \in \mathcal{B}(Y; X)$ satisfying

$$Z(\mathbf{y}) = D_1 f(g(\mathbf{y}), \mathbf{y})Z(\mathbf{y}) + D_2 f(g(\mathbf{y}), \mathbf{y}). \quad (7.7)$$

Moreover, Z is continuous by Lemma 7.2.6.

Lemma 7.2.9. *Let X, Y, U, V, f , and g be as in the hypothesis of Theorem 7.2.4. If f is C^1 at each $(g(\mathbf{y}), \mathbf{y}) \in \overline{U} \times V$, then g is C^1 at each $\mathbf{y} \in V$.*

The proof of this lemma is hard. It is given in Section 7.2.2.

We are now finally ready to prove the uniform contraction mapping principle (Theorem 7.2.4).

Proof of Theorem 7.2.4. By Lemma 7.2.9, if $f \in C^1(\overline{U} \times V; \overline{U})$, then $g \in C^1(V; \overline{U})$. We assume that the theorem holds for $n = k - 1 \in \mathbb{N}$ and prove by induction that the theorem holds for $n = k$. If f is C^k , then g is at least C^{k-1} . Since $Dg(\mathbf{y})$ satisfies (7.7), and ϕ is a C^{k-1} uniform contraction, we have that $Dg(\mathbf{y})$ is also C^{k-1} by the inductive hypothesis. Hence, g is C^k . $\square$

7.2.2 * Proof of Lemma 7.2.9

Proof. If we knew that Dg existed, then for each $\mathbf{y} \in V$ we would have $Dg(\mathbf{y}) \in \mathcal{B}(Y; X)$, and applying the chain rule to (7.5) would give

$$Dg(\mathbf{y}) = D_1 f(g(\mathbf{y}), \mathbf{y})Dg(\mathbf{y}) + D_2 f(g(\mathbf{y}), \mathbf{y}). \quad (7.8)$$

That means that $Dg(\mathbf{y})$ would be a fixed point of the function $\phi : \mathcal{B}(Y; X) \times V \rightarrow \mathcal{B}(Y; X)$ defined in Lemma 7.2.7. By Lemma 7.2.7, the map ϕ is a uniform contraction mapping, and so there exists a function $Z : V \rightarrow \mathcal{B}(Y; X)$, defined by setting $Z(\mathbf{y})$ equal to the unique fixed point of $\phi(\cdot, \mathbf{y})$. By Lemma 7.2.6 the map Z is continuous; therefore, all that remains is to show that $Z(\mathbf{y}) = Dg(\mathbf{y})$ for each $\mathbf{y}$. That is, for every $\varepsilon > 0$ we must show there exists $B(\mathbf{y}, \delta) \subset V$ such that if $\|\mathbf{k}\|_Y < \delta$, then

$$\|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y}) - Z(\mathbf{y})\mathbf{k}\|_X \leq \varepsilon \|\mathbf{k}\|_Y. \quad (7.9)$$

We now prove (7.9). For any $\mathbf{h} \in X$ and $\mathbf{k} \in Y$, let $\Delta(\mathbf{h}, \mathbf{k})$ be given by

$$\Delta(\mathbf{h}, \mathbf{k}) = f(g(\mathbf{y}) + \mathbf{h}, \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y}) - D_1 f(g(\mathbf{y}), \mathbf{y})\mathbf{h} - D_2 f(g(\mathbf{y}), \mathbf{y})\mathbf{k}. \quad (7.10)$$

If f is C^1 , then for all $\eta > 0$, there exists $B(\mathbf{y}, \delta_0) \subset V$ such that

$$\|\Delta(\mathbf{h}, \mathbf{k})\|_X \leq \eta(\|\mathbf{h}\|_X + \|\mathbf{k}\|_Y) \quad (7.11)$$

whenever $\|\mathbf{k}\|_Y < \delta_0$ and $\|\mathbf{h}\|_X < \delta_0$ is such that $g(\mathbf{y}) + \mathbf{h} \in \overline{U}$.

The expression to control is $\|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y}) - Z(\mathbf{y})\mathbf{k}\|_X$, so to connect this to (7.11), we choose a very specific form for $\mathbf{h}$ as a function of $\mathbf{k}$ by setting $\mathbf{h}(\mathbf{k}) = g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y})$. Note that since $g(\mathbf{y} + \mathbf{k}) \in \overline{U}$, this choice of $\mathbf{h}(\mathbf{k})$ satisfies the

requirement $g(\mathbf{y}) + \mathbf{h}(\mathbf{k}) \in \overline{U}$ for (7.11) to hold. Since g is continuous, the function $\mathbf{h}$ is continuous, and thus there exists $B(\mathbf{y}, \delta) \subset B(\mathbf{y}, \delta_0) \subset V$ such that $\|\mathbf{h}(\mathbf{k})\|_X \leq \delta_0$ whenever $\|\mathbf{k}\|_Y < \delta$. Moreover, we have

$$\begin{aligned}\mathbf{h}(\mathbf{k}) &= f(g(\mathbf{y} + \mathbf{k}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y}) \\ &= f(g(\mathbf{y}) + \mathbf{h}(\mathbf{k}), \mathbf{y} + \mathbf{k}) - f(g(\mathbf{y}), \mathbf{y}).\end{aligned}$$

Combining this with (7.10), we have

$$\mathbf{h}(\mathbf{k}) = D_1 f(g(\mathbf{y}), \mathbf{y})\mathbf{h}(\mathbf{k}) + D_2 f(g(\mathbf{y}), \mathbf{y})\mathbf{k} + \Delta(\mathbf{h}(\mathbf{k}), \mathbf{k}),$$

which yields

$$\|\mathbf{h}(\mathbf{k})\|_X \leq \|D_1 f(g(\mathbf{y}), \mathbf{y})\| \|\mathbf{h}(\mathbf{k})\|_X + \|D_2 f(g(\mathbf{y}), \mathbf{y})\| \|\mathbf{k}\|_Y + \eta(\|\mathbf{h}(\mathbf{k})\|_X + \|\mathbf{k}\|_Y).$$

Choose $\eta \leq \frac{1-\lambda}{2}$, and recall that $\|D_1 f(g(\mathbf{y}), \mathbf{y})\| \leq \lambda$. Combining these, we find

$$\|\mathbf{h}(\mathbf{k})\|_X \leq \frac{\|D_2 f(g(\mathbf{y}), \mathbf{y})\| + \eta}{1 - \lambda - \eta} \|\mathbf{k}\|_Y \leq \underbrace{\frac{2\|D_2 f(g(\mathbf{y}), \mathbf{y})\| + 1 - \lambda}{1 - \lambda}}_M \|\mathbf{k}\|_Y.$$

Setting $M = \frac{2\|D_2 f(g(\mathbf{y}), \mathbf{y})\| + 1 - \lambda}{1 - \lambda}$ and simplifying (7.11) yields

$$\|\Delta(\mathbf{h}(\mathbf{k}), \mathbf{k})\|_X < \eta(M + 1)\|\mathbf{k}\|_Y \quad (7.12)$$

whenever $\|\mathbf{k}\|_Y \leq \delta$. Moreover from (7.7), we have

$$\begin{aligned}\Delta(\mathbf{h}(\mathbf{k}), \mathbf{k}) &= \mathbf{h}(\mathbf{k}) - D_1 f(g(\mathbf{y}), \mathbf{y})\mathbf{h}(\mathbf{k}) - D_2 f(g(\mathbf{y}), \mathbf{y})\mathbf{k} \\ &= \mathbf{h}(\mathbf{k}) - D_1 f(g(\mathbf{y}), \mathbf{y})\mathbf{h}(\mathbf{k}) - Z(\mathbf{y})\mathbf{k} + D_1 f(g(\mathbf{y}), \mathbf{y})Z(\mathbf{y})\mathbf{k} \\ &= (I - D_1 f(g(\mathbf{y}), \mathbf{y}))(\mathbf{h}(\mathbf{k}) - Z(\mathbf{y})\mathbf{k}).\end{aligned}$$

Since $\|D_1 f(g(\mathbf{y}), \mathbf{y})\| \leq \lambda < 1$, we have that $I - D_1 f(g(\mathbf{y}), \mathbf{y})$ is invertible (see Proposition 5.7.4). This, combined with (7.12), yields

$$\begin{aligned}\|g(\mathbf{y} + \mathbf{k}) - g(\mathbf{y}) - Z(\mathbf{y})\mathbf{k}\|_X &= \|\mathbf{h}(\mathbf{k}) - Z(\mathbf{y})\mathbf{k}\|_X \\ &= \|(I - D_1 f(g(\mathbf{y}), \mathbf{y}))^{-1} \Delta(\mathbf{h}(\mathbf{k}), \mathbf{k})\|_X \\ &\leq \|(I - D_1 f(g(\mathbf{y}), \mathbf{y}))^{-1}\| \|\Delta(\mathbf{h}(\mathbf{k}), \mathbf{k})\|_X \\ &< \frac{\eta(M + 1)}{1 - \lambda} \|\mathbf{k}\|_Y\end{aligned}$$

whenever $\|\mathbf{k}\|_Y < \delta$. Finally, setting $\eta = \min\{\frac{\epsilon(1-\lambda)}{M+1}, \frac{1-\lambda}{2}\}$ gives (7.9). $\square$

7.3 Newton's Method

Finding zeros of a function is an essential step for solving many problems in both pure and applied mathematics. Some of the fastest and most widely used methods for solving these sorts of problems are Newton's method and its variants. Newton's method is a fundamental tool in many important algorithms; indeed, even many

algorithms that are not obviously built from Newton's method really have Newton's method at their core, once you look carefully [HRW12, Tap09].

The idea of Newton's method is simple: if finding a zero of a function is difficult, replace the function with a simpler approximation whose zeros are easier to find. Zeros of linear functions are easy to find, so the obvious choice is a linear approximation. Given a differentiable function $f : X \rightarrow X$, the best linear approximation to f at $\mathbf{x}_n$ is the function

$$L(\mathbf{x}) = f(\mathbf{x}_n) + Df(\mathbf{x}_n)(\mathbf{x} - \mathbf{x}_n).$$

Assuming that $Df(\mathbf{x}_n)$ is invertible, this linear system has a unique zero at

$$\mathbf{x}_{n+1} = \mathbf{x}_n - Df(\mathbf{x}_n)^{-1}f(\mathbf{x}_n). \quad (7.13)$$

So $\mathbf{x}_{n+1}$ should be a better approximation of a zero of f than $\mathbf{x}_n$ was. Starting at any $\mathbf{x}_0$ and repeating for each $n \in \mathbb{N}$ gives a sequence $\mathbf{x}_0, \mathbf{x}_1, \dots$ that often (but not always) converges to a zero of f . Moreover, if $\mathbf{x}_0$ is chosen well, the convergence can be very fast. See Figure 7.1 for an illustration.

In this section we give some conditions that guarantee both that the method converges and that the convergence is rapid. We first prove some of these results in one dimension, and then we generalize to higher dimensions and even Banach spaces.

7.3.1 Convergence

Before treating Newton's method and its variants, we need to make a brief digression into convergence rates. These are discussed much more in Volume 2.

An iterative process produces a sequence $(\mathbf{x}_n)_{n=0}^\infty$ of approximations. If we are approximating $\bar{\mathbf{x}}$, we expect the sequence to converge to $\bar{\mathbf{x}}$. Better algorithms produce a sequence that converges more rapidly.

Definition 7.3.1. *Given a sequence $(\mathbf{x}_n)_{n=0}^\infty$ approximating $\bar{\mathbf{x}}$, denote the error of the n th approximation by*

$$\varepsilon_n = \|\mathbf{x}_n - \bar{\mathbf{x}}\|.$$

The sequence is said to converge linearly with rate $\mu < 1$ if

$$\varepsilon_{n+1} \leq \mu \varepsilon_n$$

for each $n \in \mathbb{N}$. The sequence is said to converge quadratically, when there exists a constant $k \geq 0$ (not necessarily less than 1) such that

$$\varepsilon_{n+1} \leq k \varepsilon_n^2$$

for all $n \in \mathbb{N}$.

If a sequence of real numbers converges linearly with rate μ , then with each iteration the approximation adds about $-\log_{10} \mu$ digits of accuracy. Quadratic convergence means that with each iteration the approximation roughly doubles the number of digits of accuracy. This is much better than linear convergence.

The quasi-Newton method of Section 7.3.3 converges linearly. In Theorems 7.3.4 and 7.3.12 we show that Newton's method converges quadratically.

7.3.2 Newton's Method: Scalar Version

To prove that Newton's method converges, we construct a contraction mapping.

Lemma 7.3.2. *Let $f : [a, b] \rightarrow \mathbb{R}$ be C^2 and assume that for some $\bar{x} \in (a, b)$ we have $f(\bar{x}) = 0$ and $f'(\bar{x}) \neq 0$. Under these hypotheses there exists $\delta > 0$ such that the map*

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

is a contraction on $[\bar{x} - \delta, \bar{x} + \delta] \subset [a, b]$.

Proof. Choose $\delta > 0$ so that $[\bar{x} - \delta, \bar{x} + \delta] \subset [a, b]$. Because f is C^2 we can shrink δ so that for all $x \in [\bar{x} - \delta, \bar{x} + \delta] \subset [a, b]$ we have $f'(x) \neq 0$ and

$$k = \sup_{x \in [\bar{x} - \delta, \bar{x} + \delta]} \left| \frac{f(x)f''(x)}{f'(x)^2} \right| < 1.$$

By the mean value theorem, for any $[x, y] \subset [\bar{x} - \delta, \bar{x} + \delta]$ there exists $c \in [x, y]$ such that

$$|\phi(x) - \phi(y)| = |\phi'(c)||x - y| = \left| \frac{f(c)f''(c)}{f'(c)^2} \right| |x - y| \leq k|x - y|.$$

Therefore, ϕ is a contraction on $[\bar{x} - \delta, \bar{x} + \delta]$. $\square$

Remark 7.3.3. We know that $\phi([\bar{x} - \delta, \bar{x} + \delta]) \subset [\bar{x} - \delta, \bar{x} + \delta]$ because ϕ is a contraction and $\bar{x}$ is a fixed point. Thus,

$$|\phi(\bar{x} \pm \delta) - \bar{x}| = |\phi(\bar{x} \pm \delta) - \phi(\bar{x})| \leq k|\bar{x} \pm \delta - \bar{x}| = k\delta < \delta.$$

Theorem 7.3.4 (Newton's Method—Scalar Version). *If $f : [a, b] \rightarrow \mathbb{R}$ satisfies the hypotheses of the lemma above, then the iterative map*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{for all } n \in \mathbb{N} \tag{7.14}$$

converges to $\bar{x}$ quadratically, whenever x_0 is sufficiently close to $\bar{x}$.

Proof. Since f is C^2 the derivative f' is locally Lipschitz at the point $\bar{x}$ with some constant L (see Proposition 6.3.7), so there exists a δ_1 such that $|f'(\bar{x} + \varepsilon) - f'(\bar{x})| < L|\varepsilon|$ whenever $|\varepsilon| < \delta_1$. Let $\delta < \delta_1$ be chosen as in the previous lemma. Choose any initial $x_0 \in [\bar{x} - \delta, \bar{x} + \delta]$ and iterate. By the lemma, the sequence must converge to $\bar{x}$.

Let $\varepsilon_n = x_n - \bar{x}$ for each $n \in \mathbb{N}$. By the mean value theorem $f(\bar{x} + \varepsilon_{n-1}) = f(\bar{x}) + f'(\bar{x} + \eta\varepsilon_{n-1})\varepsilon_{n-1}$ for some $\eta \in [0, 1]$ (convince yourself that this still holds

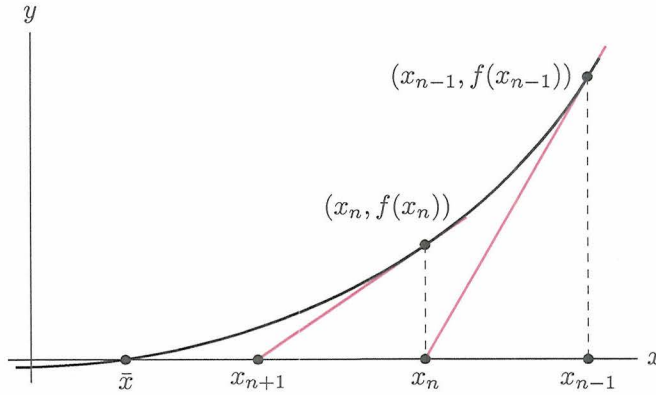


Figure 7.1. Newton's method takes the tangent line (red) to the curve $y = f(x)$ at the point $(x_n, f(x_n))$ and defines x_{n+1} to be the x -intercept of that line. Details are given in Theorem 7.3.4.

if $\varepsilon_{n-1} < 0$). Thus, from (7.14) we have

$$\begin{aligned}
 |\varepsilon_n| &= \left| \varepsilon_{n-1} - \frac{f(\bar{x} + \varepsilon_{n-1})}{f'(\bar{x} + \varepsilon_{n-1})} \right| \\
 &= \left| \frac{f'(\bar{x} + \varepsilon_{n-1})\varepsilon_{n-1} - f(\bar{x} + \varepsilon_{n-1})}{f'(\bar{x} + \varepsilon_{n-1})} \right| \\
 &\leq \left| \frac{f'(\bar{x} + \varepsilon_{n-1}) - f'(\bar{x} + \eta\varepsilon_{n-1})}{f'(\bar{x} + \varepsilon_{n-1})} \right| |\varepsilon_{n-1}| \\
 &\leq \frac{L(1 - \eta)|\varepsilon_{n-1}|}{|f'(\bar{x} + \varepsilon_{n-1})|} |\varepsilon_{n-1}| \\
 &\leq M|\varepsilon_{n-1}|^2,
 \end{aligned}$$

where $M = \sup_{x \in [\bar{x} - \delta, \bar{x} + \delta]} \frac{L}{|f'(x)|}$. $\square$

Example 7.3.5. Applying Newton's method on the function $f(x) = x^2 - a$ gives the following fast algorithm for computing $\sqrt{a}$:

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right). \quad (7.15)$$

This is derived in Exercise 7.13. Using $x_0 = 1.0$ and $a = 4$ we get 14 digits of accuracy in five iterations:

$$\begin{aligned}
 x_1 &= 2.5000000000000000, \\
 x_2 &= 2.0500000000000000,
 \end{aligned}$$

$$x_3 = 2.000609756097561,$$

$$x_4 = 2.000000092922295,$$

$$x_5 = 2.000000000000002.$$

Notice how quickly this sequence converges to 2.

Remark 7.3.6. If the derivative of f vanishes at $\bar{x}$, then Newton's method is not necessarily quadratic and may not even converge at all! If $f'(\bar{x}) = 0$, we say that f has a *multiple zero* at $\bar{x}$. For example, if f is a polynomial and $f'(\bar{x}) = 0$, then f has a factor (over $\mathbb{C}$) of the form $(x - \bar{x})^2$.

Example 7.3.7.* Using Newton's method on the function $f(x) = x^3 - a$ gives the following fast algorithm for computing $\sqrt[3]{a}$:

$$x_n = x_{n-1} + \frac{a - x_{n-1}^3}{3x_{n-1}^2}. \quad (7.16)$$

Using $a = 1729.03$ and $x_0 = 12$, we get 15 digits of accuracy after three iterations:

$$x_1 = 12.002384259259259,$$

$$x_2 = 12.002383785691737,$$

$$x_3 = 12.002383785691718.$$

In Richard Feynman's book *Surely You're Joking, Mr. Feynman!* [FLH85], he tells a story of an abacus master who challenges him to a race to solve various arithmetic problems. The man was using his abacus, and Feynman was using pen and paper. After easily beating Feynman in various multiplication problems, the abacus master challenged him to a division problem, which turned out a tie. Frustrated that he had not won the division contest, the abacus master challenged Feynman to find the cube root of 1729.03. This was a mistake, because computing cube roots on an abacus is hard work, but Feynman was an expert in algorithms, and Newton's method was in his arsenal. He also knew that $12^3 = 1728$ since there are 1728 cubic inches in a cubic foot. Then using (7.16), he carried out the following estimate

$$\sqrt[3]{1729.03} \approx 12 + \frac{1729.03 - 1728}{3 \cdot 12^2} = 12 + \frac{1.03}{432} \approx 12.002. \quad (7.17)$$

Feynman won this last contest easily, finding the answer to three decimal places before the abacus master could find one.

The algorithm (7.16) would have allowed Feynman to compute the cube root to 5 decimals of accuracy in a single iteration had he computed the

fraction in (7.17) to more decimal places instead of his quick estimation. In three iterations, he would have been able to get to 15 digits of accuracy. That would have taken forever with an abacus!

Remark 7.3.8. It is essential in Newton's method (7.14) that the initial point x_0 be sufficiently close to the zero. If it is not close enough, it is possible that the sequence will bounce around and never converge, or even go to infinity.

Unexample 7.3.9. Using Newton's method on the function $f(x) = x^{1/3}$, we have

$$x_n = x_{n-1} + \frac{-x_{n-1}^{1/3}}{\frac{1}{3}x_{n-1}^{-2/3}} = -2x_{n-1}.$$

The initial guess $x_0 = 1$ gives the sequence $1, -2, 4, -8, 16, \dots$. Clearly this sequence does not converge to the zero of $x^{1/3}$.

7.3.3 A Quasi-Newton Method: Vector Version

We now describe a quasi-Newton method for vector-valued functions that does not converge as quickly as Newton's method but is easier to prove converges. This method also plays an important role in the proof of the implicit function theorem in the next section.

Theorem 7.3.10. *Let $(X, \|\cdot\|)$ be a Banach space and assume $f : X \rightarrow X$ is C^1 on an open neighborhood U of the point $\bar{\mathbf{x}}$. If $f(\bar{\mathbf{x}}) = \mathbf{0}$ and $Df(\bar{\mathbf{x}}) \in \mathcal{B}(X)$ has a bounded inverse, then there exists $\delta > 0$ such that*

$$\phi(\mathbf{x}) = \mathbf{x} - Df(\bar{\mathbf{x}})^{-1}f(\mathbf{x}) \quad (7.18)$$

is a contraction on $\overline{B(\bar{\mathbf{x}}, \delta)}$.

Proof. Since $Df(\mathbf{x})$ is continuous on U , we can choose $\overline{B(\bar{\mathbf{x}}, \delta)} \subset U$ so that

$$\|Df(\bar{\mathbf{x}}) - Df(\mathbf{x})\| < \frac{1}{2\|Df(\bar{\mathbf{x}})^{-1}\|},$$

whenever $\|\bar{\mathbf{x}} - \mathbf{x}\| < \delta$. Hence, $\mathbf{x} \in \overline{B(\bar{\mathbf{x}}, \delta)}$ implies

$$\|D\phi(\mathbf{x})\| = \|I - Df(\bar{\mathbf{x}})^{-1}Df(\mathbf{x})\| = \|Df(\bar{\mathbf{x}})^{-1}\| \|Df(\bar{\mathbf{x}}) - Df(\mathbf{x})\| < \frac{1}{2}.$$

Thus, by the mean value theorem we have $\|\phi(\mathbf{x}) - \phi(\mathbf{y})\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \overline{B(\bar{\mathbf{x}}, \delta)}$. $\square$

The previous theorem does not immediately give us an algorithm because it depends on computing $Df(\bar{\mathbf{x}})$, which we generally do not know unless we know $\bar{\mathbf{x}}$. Many quasi-Newton methods amount to choosing a suitable approximation to

$Df(\bar{\mathbf{x}})^{-1}$, which we can then use in the contraction mapping above to produce an iterative algorithm.

The following lemma provides a useful tool for approximating $Df(\bar{\mathbf{x}})^{-1}$ and is also important for proving convergence of Newton's method.

Lemma 7.3.11. *Let $(X, \|\cdot\|)$ be a Banach space and assume that $g : X \rightarrow \mathcal{B}(X)$ is a continuous map. If $g(\bar{\mathbf{x}})$ has a bounded inverse, then there exists $\delta > 0$ such that $\|g(\mathbf{x})^{-1}\| < 2\|g(\bar{\mathbf{x}})^{-1}\|$ whenever $\|\bar{\mathbf{x}} - \mathbf{x}\| < \delta$.*

Proof. Since g and matrix inversion are both continuous in a neighborhood of $\bar{\mathbf{x}}$ (see Proposition 5.7.7), so is their composition. Hence, if $\varepsilon = \|g(\bar{\mathbf{x}})^{-1}\|$, then there exists $\delta > 0$ such that $\|g(\mathbf{x})^{-1} - g(\bar{\mathbf{x}})^{-1}\| < \varepsilon$ whenever $\|\mathbf{x} - \bar{\mathbf{x}}\| < \delta$. Thus,

$$\|g(\mathbf{x})^{-1}\| \leq \|g(\mathbf{x})^{-1} - g(\bar{\mathbf{x}})^{-1}\| + \|g(\bar{\mathbf{x}})^{-1}\| < 2\|g(\bar{\mathbf{x}})^{-1}\|$$

whenever $\|\mathbf{x} - \bar{\mathbf{x}}\| < \delta$. $\square$

7.3.4 Newton's Method: Vector Version

The idea of Newton's method in general is the same as in the single-variable case. The derivative at a point provides the best possible linear approximation to the function near that point, and if the derivative is invertible, we can find the zero of this linear approximation and use it as the next iterate, namely, (7.19) below.

The hard part of all this is proving that convergence is quadratic for this method. But even here the idea is similar to the single-variable case, that is, to examine the remainder R_2 in the Taylor polynomial of f .

Theorem 7.3.12 (Newton's Method—Vector Version). *Let $(X, \|\cdot\|)$ be a Banach space and assume $f : X \rightarrow X$ is C^1 in an open neighborhood U of the point $\bar{\mathbf{x}} \in X$ and $f(\bar{\mathbf{x}}) = 0$. If $Df(\bar{\mathbf{x}}) \in \mathcal{B}(X)$ has a bounded inverse and $Df(\mathbf{x})$ is Lipschitz on U , then the iterative map*

$$\mathbf{x}_{n+1} = \mathbf{x}_n - Df(\mathbf{x}_n)^{-1}f(\mathbf{x}_n) \tag{7.19}$$

converges quadratically to $\bar{\mathbf{x}}$ whenever $\mathbf{x}_0$ is sufficiently close to $\bar{\mathbf{x}}$.

Proof. Choose $\delta > 0$ as in Lemma 7.3.11, and such that $B(\bar{\mathbf{x}}, \delta) \subset U$. Let $\mathbf{x}_0 \in B(\bar{\mathbf{x}}, \delta)$, and define $\mathbf{x}_n$ for $n > 0$ using (7.19). We begin by writing the integral remainder of the first-order Taylor expansion

$$\begin{aligned} f(\mathbf{x}_n) - f(\bar{\mathbf{x}}) &= \int_0^1 Df(\bar{\mathbf{x}} + t(\mathbf{x}_n - \bar{\mathbf{x}}))(\mathbf{x}_n - \bar{\mathbf{x}}) dt \\ &= Df(\bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}}) + \int_0^1 (Df(\bar{\mathbf{x}} + t(\mathbf{x}_n - \bar{\mathbf{x}})) - Df(\bar{\mathbf{x}}))(\mathbf{x}_n - \bar{\mathbf{x}}) dt. \end{aligned}$$

Assume that k is the Lipschitz constant for Df on U . By the previous line, we have

$$\begin{aligned} \|f(\mathbf{x}_n) - f(\bar{\mathbf{x}}) - Df(\bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})\| &\leq \int_0^1 \|Df(\bar{\mathbf{x}} + t(\mathbf{x}_n - \bar{\mathbf{x}})) - Df(\bar{\mathbf{x}})\| \|\mathbf{x}_n - \bar{\mathbf{x}}\| dt \\ &\leq \int_0^1 k \|\bar{\mathbf{x}} + t(\mathbf{x}_n - \bar{\mathbf{x}}) - \bar{\mathbf{x}}\| \|\mathbf{x}_n - \bar{\mathbf{x}}\| dt \\ &= \int_0^1 kt \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 dt = \frac{k}{2} \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2. \end{aligned}$$

Also, from (7.19), we have

$$\begin{aligned} \mathbf{x}_{n+1} - \bar{\mathbf{x}} &= \mathbf{x}_n - Df(\mathbf{x}_n)^{-1} f(\mathbf{x}_n) - \bar{\mathbf{x}} + Df(\mathbf{x}_n)^{-1} f(\bar{\mathbf{x}}) \\ &= \mathbf{x}_n - \bar{\mathbf{x}} - Df(\mathbf{x}_n)^{-1} (f(\mathbf{x}_n) - f(\bar{\mathbf{x}})) \\ &= \mathbf{x}_n - \bar{\mathbf{x}} - Df(\mathbf{x}_n)^{-1} (Df(\bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}}) + f(\mathbf{x}_n) - f(\bar{\mathbf{x}}) - Df(\bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})) \\ &= Df(\mathbf{x}_n)^{-1} (Df(\mathbf{x}_n) - Df(\bar{\mathbf{x}}))(\mathbf{x}_n - \bar{\mathbf{x}}) \\ &\quad - Df(\mathbf{x}_n)^{-1} (f(\mathbf{x}_n) - f(\bar{\mathbf{x}}) - Df(\bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})). \end{aligned}$$

Thus, for $M = 3k\|Df(\bar{\mathbf{x}})^{-1}\|$, we have

$$\begin{aligned} \|\mathbf{x}_{n+1} - \bar{\mathbf{x}}\| &\leq \|Df(\mathbf{x}_n)^{-1}\| \|Df(\mathbf{x}_n) - Df(\bar{\mathbf{x}})\| \|\mathbf{x}_n - \bar{\mathbf{x}}\| + \frac{k}{2} \|Df(\mathbf{x}_n)^{-1}\| \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 \\ &\leq \frac{3}{2} k \|Df(\mathbf{x}_n)^{-1}\| \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 \\ &\leq 3k \|Df(\bar{\mathbf{x}})^{-1}\| \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 \\ &\leq M \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2. \quad \square \end{aligned}$$

Remark 7.3.13. In the previous theorem, we proved that when the initial point $\mathbf{x}_0$ is “sufficiently close” to the zero $\bar{\mathbf{x}}$, then Newton’s method converges. But to be useful, we also need to know whether a given starting point will converge. This is answered by the Newton–Kantorovich theorem, which is a generalization of Lemma 7.3.2 to vector-valued functions. It says that the initial value $\mathbf{x}_0$ produces a convergent sequence if

$$\|f(\mathbf{x}_0)\| \|Df(\mathbf{x}_0)^{-1}\|^2 K \leq \frac{1}{2}. \quad (7.20)$$

Here K is the Lipschitz constant for the map $Df : U \rightarrow \mathcal{B}(X)$. The proof of the Newton–Kantorovich theorem is not beyond the scope of the text, but it is tedious, so we do not reproduce it here.

Example 7.3.14. We apply Newton’s method to find the zeros of the function

$$f(x, y) = [4x^3 + 4xy^2 \quad 4x^2y + 4y^3]^\top.$$

Observe that

$$Df(\mathbf{x}) = \begin{bmatrix} 12x^2 + 4y^2 & 8xy \\ 8xy & 4x^2 + 12y^2 \end{bmatrix},$$

and since $Df(\mathbf{x})$ is a 2×2 matrix we calculate $Df(\mathbf{x})^{-1}$ directly as

$$Df(\mathbf{x})^{-1} = \frac{1}{12(x^2 + y^2)^2} \begin{bmatrix} x^2 + 3y^2 & -2xy \\ -2xy & 3x^2 + y^2 \end{bmatrix}.$$

Suppose we select $\begin{bmatrix} a & a \end{bmatrix}^T$ as our initial guess where $a \neq 0$. Then

$$\mathbf{x}_1 = \mathbf{x}_0 - Df(\mathbf{x}_0)^{-1}f(\mathbf{x}_0) = \begin{bmatrix} a \\ a \end{bmatrix} - \frac{1}{48a^4} \begin{bmatrix} 4a^2 & -2a^2 \\ -2a^2 & 4a^2 \end{bmatrix} \begin{bmatrix} 8a^3 \\ 8a^3 \end{bmatrix} = \begin{bmatrix} 2a/3 \\ 2a/3 \end{bmatrix} = \frac{2}{3} \begin{bmatrix} a \\ a \end{bmatrix}.$$

It is clear that in general $\mathbf{x}_n = \left(\frac{2}{3}\right)^n \begin{bmatrix} a \\ a \end{bmatrix}$, which converges quickly to the zero at $(0, 0)$.

7.4 The Implicit and Inverse Function Theorems

The implicit function theorem and the inverse function theorem are consequences of the uniform contraction mapping principle and are the basis of many ideas in mathematics. Although most analysis texts prove these only in $\mathbb{R}^n$, we can do much better and prove them in a general Banach space with essentially no extra work.

7.4.1 Implicit Function Theorem

Given a function of two variables $F : X \times Y \rightarrow Z$, each point $\mathbf{z}_0 \in Z$ defines a level set $\{(\mathbf{x}, \mathbf{y}) | F(\mathbf{x}, \mathbf{y}) = \mathbf{z}_0\}$. Globally, this set is not usually the graph of a single function, but the implicit function theorem says that under mild conditions the level set is *locally* the graph of a function—call this function f . This new function f is the *implicit function* defined by the relation $F(\mathbf{x}, \mathbf{y}) = \mathbf{z}_0$ near some point $(\mathbf{x}_0, \mathbf{y}_0)$.

Example 7.4.1. Consider the level set $\{F(x, y) = 9\}$ of the function $F(x, y) = x^2 + y^2$. In a neighborhood of the point $(x_0, y_0) = (0, 3)$ we can define y as a function of x , namely, $y = \sqrt{9 - x^2}$. However, we cannot define y as a function of x in a neighborhood around the point $(3, 0)$ since in any neighborhood of $(3, 0)$ there are two points of the level set of the form $(x, \pm\sqrt{9 - x^2})$ with the same x -coordinate. This is depicted in Figure 7.2

The implicit function theorem tells us when we can implicitly define one or more of the variables (in our example the variable y) as functions of other variables (in our example the variable x).

In the previous example, we could solve explicitly for y as a function of x , but in many cases solving explicitly for the function is just too hard (even impossible). Yet,

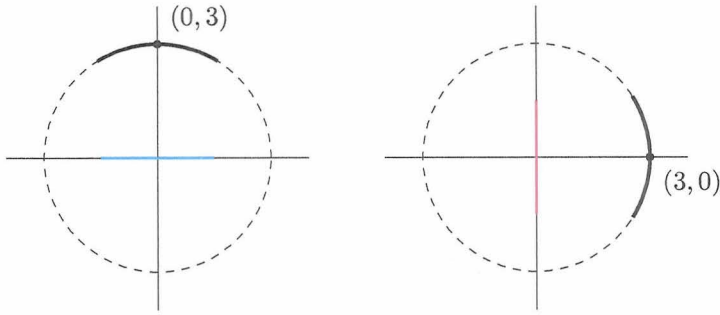


Figure 7.2. An illustration of Example 7.4.1. In a neighborhood around the point $(0,3)$, the points on the circle $F(x,y) = 9$ (black arc on the left) can be written as $(x, f(x))$, provided x remains in a small enough neighborhood (blue line) of 0. But near the point $(3,0)$ there is no function of x defining y . Instead, there is a function g so that we can write points of the circle near $(3,0)$ as $(g(y), y)$, provided y remains in a small enough neighborhood (red line) of 0.

for many problems just knowing it exists and knowing its derivative is enough. The implicit function theorem not only tells us when the function exists, but also how to compute its derivative without computing the function itself; see (7.22).

To prove the implicit function theorem we construct a uniform contraction mapping using a generalization of the quasi-Newton method (7.18) and then apply the uniform contraction mapping principle (Theorem 7.2.4).

Theorem 7.4.2 (Implicit Function Theorem). Assume that $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, and $(Z, \|\cdot\|_Z)$ are Banach spaces, that U and V are open neighborhoods of $\mathbf{x}_0 \in X$ and $\mathbf{y}_0 \in Y$, respectively, and that $F : U \times V \rightarrow Z$ is a C^k map for some integer $k \geq 1$. Let $\mathbf{z}_0 = F(\mathbf{x}_0, \mathbf{y}_0)$. If $D_2F(\mathbf{x}_0, \mathbf{y}_0) \in \mathcal{B}(Y; Z)$ has a bounded inverse, then there exists an open neighborhood $U_0 \times V_0 \subset U \times V$ of $(\mathbf{x}_0, \mathbf{y}_0)$ and a unique C^k function $f : U_0 \rightarrow V_0$ such that $f(\mathbf{x}_0) = \mathbf{y}_0$ and

$$\{(\mathbf{x}, \mathbf{y}) \in U_0 \times V_0 \mid F(\mathbf{x}, \mathbf{y}) = \mathbf{z}_0\} = \{(\mathbf{x}, f(\mathbf{x})) \mid \mathbf{x} \in U_0\}. \quad (7.21)$$

Moreover, the derivative of f satisfies

$$Df(\mathbf{x}) = -D_2F(\mathbf{x}, f(\mathbf{x}))^{-1}D_1F(\mathbf{x}, f(\mathbf{x})) \quad (7.22)$$

on U_0 .

Proof. Without loss of generality, we assume that $F(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ (otherwise redefine F to be $F - \mathbf{z}_0$). Generalizing (7.18), consider the C^k map $G : U \times V \rightarrow Y$ given by

$$G(\mathbf{x}, \mathbf{y}) = \mathbf{y} - D_2F(\mathbf{x}_0, \mathbf{y}_0)^{-1}F(\mathbf{x}, \mathbf{y}).$$

For fixed $\mathbf{x} \in U$, we have that $G(\mathbf{x}, \mathbf{y}) = \mathbf{y}$ if and only if $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. Furthermore, $D_2G(\mathbf{x}_0, \mathbf{y}_0) = I - D_2F(\mathbf{x}_0, \mathbf{y}_0)^{-1}D_2F(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$. Since G is C^1 , there exists a neighborhood $U_1 \times \overline{V}_0 \subset U \times V$ of $(\mathbf{x}_0, \mathbf{y}_0)$ such that

$$\|D_2G(\mathbf{x}, \mathbf{y})\| < \frac{1}{2}$$

whenever $(\mathbf{x}, \mathbf{y}) \in U_1 \times \overline{V_0}$. Without loss of generality we may assume $V_0 = B(\mathbf{y}_0, \delta)$ for some suitably chosen $\delta > 0$. Since $F(\mathbf{x}, \mathbf{y})$ is C^1 and vanishes at $(\mathbf{x}_0, \mathbf{y}_0)$, there exists an open neighborhood $U_0 \subset U_1$ of $\mathbf{x}_0$ such that

$$\|D_2F(\mathbf{x}_0, \mathbf{y}_0)^{-1}\| \|F(\mathbf{x}, \mathbf{y}_0)\| < \frac{\delta}{2}$$

whenever $\mathbf{x} \in U_0$. Applying the triangle inequality and the inequality (6.18) from the integral mean value theorem (Corollary 6.5.5), we have

$$\begin{aligned} \|G(\mathbf{x}, \mathbf{y}) - \mathbf{y}_0\| &\leq \|G(\mathbf{x}, \mathbf{y}) - G(\mathbf{x}, \mathbf{y}_0)\| + \|G(\mathbf{x}, \mathbf{y}_0) - \mathbf{y}_0\| \\ &\leq \sup_{\tilde{\mathbf{y}} \in \ell(\mathbf{y}_0, \mathbf{y})} \|D_2G(\mathbf{x}, \tilde{\mathbf{y}})\| \|\mathbf{y} - \mathbf{y}_0\| + \|D_2F(\mathbf{x}_0, \mathbf{y}_0)^{-1}\| \|F(\mathbf{x}, \mathbf{y}_0)\| \\ &< \frac{\delta}{2} + \frac{\delta}{2} = \delta \end{aligned}$$

whenever $(\mathbf{x}, \mathbf{y}) \in U_0 \times \overline{V_0}$, and thus $G : U_0 \times \overline{V_0} \rightarrow \overline{V_0}$. Moreover, for $\mathbf{x} \in U_0$ and $\mathbf{y}_1, \mathbf{y}_2 \in \overline{V_0}$, we apply the mean value inequality (6.18) again to get

$$\|G(\mathbf{x}, \mathbf{y}_1) - G(\mathbf{x}, \mathbf{y}_2)\| \leq \sup_{\lambda \in [0,1]} \|D_2G(\mathbf{x}, \lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2)\| \|\mathbf{y}_1 - \mathbf{y}_2\| < \frac{1}{2} \|\mathbf{y}_1 - \mathbf{y}_2\|.$$

This implies that $G(\mathbf{x}, \cdot)$ is a uniform contraction, so for each $\mathbf{x}$ there is a unique $\mathbf{y}$ satisfying $G(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. By Theorem 7.2.4, this defines a C^k function $f : U_0 \rightarrow \overline{V_0}$ satisfying $G(\mathbf{x}, f(\mathbf{x})) = f(\mathbf{x})$ for all $\mathbf{x} \in U_0$. Since $\|G(\mathbf{x}, \mathbf{y}) - \mathbf{y}_0\| < \delta$ on $U_0 \times \overline{V_0}$, we can restrict the codomain to V_0 and simply write $f : U_0 \rightarrow V_0$. It follows that $F(\mathbf{x}, f(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in U_0$, which, together with uniqueness, gives (7.21). Differentiating and solving for $Df(\mathbf{x})$ gives (7.22). $\square$

Example 7.4.3. The level set in Example 7.4.1 is a circle of radius 3, centered at the origin. By the implicit function theorem, as long as $D_2F(x_0, y_0) = 2y_0 \neq 0$, there exists a unique C^1 function $f(x)$ in a neighborhood of the point (x_0, y_0) satisfying $F(x, f(x)) = 0$.

Setting $y = f(x)$ and differentiating the equation $F(x, y) = 0$ with respect to x gives

$$0 = D_1F(x, f(x)) + D_2F(x, f(x))f'(x) = 2x + 2yf'.$$

Solving for $y' = f'(x)$ yields

$$y' = f'(x) = -\frac{D_1F(x, y)}{D_2F(x, y)} = -\frac{2x}{2y} = -\frac{x}{y},$$

which agrees with (7.22).

Notice that the tangent line to the circle at (x, y) is perpendicular to the radius connecting the origin $(0, 0)$ to (x, y) , so the slope $m = y/x$ of that radius is the negative reciprocal of the slope of the tangent line.

Remark 7.4.4. The previous example is a special case of a claim often seen in a multivariable calculus class. For any function $F(x, y)$ of two variables, if the equation $F(x, y) = 0$ defines y implicitly as a function of x , then the derivative dy/dx is given by

$$\frac{dy}{dx} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}}. \quad (7.23)$$

The implicit function theorem tells us that y is a function of x when $\frac{\partial F}{\partial y} \neq 0$, and (7.23) is a special case of the formula (7.22).

Example 7.4.5. Consider the two-dimensional surface S defined implicitly by the equation $F(x, y, z) = 0$, where

$$F(x, y, z) = z^3 + 3xyz^2 - 5x^2y^2z + 14.$$

Given $(x_0, y_0, z_0) = (1, -1, 2) \in S$, we compute $D_3F(x_0, y_0, z_0) = -5 \neq 0$. By the implicit function theorem, the surface S can be written explicitly as the graph of a function $z = z(x, y)$ in a neighborhood of (x_0, y_0, z_0) .

Furthermore, we can find the partial derivatives of z by differentiating $F(x, y, z(x, y)) = 0$, which gives

$$\begin{aligned} 0 &= D_1F(x, y, z) + D_3F(x, y, z)D_1z(x, y) \\ &= \frac{\partial F}{\partial x} + \frac{\partial F}{\partial z} \frac{\partial z}{\partial x} \\ &= (3yz^2 - 10xy^2z) + (3z^2 + 6xyz - 5x^2y^2) \frac{\partial z}{\partial x}, \\ 0 &= D_2F(x, y, z) + D_3F(x, y, z)D_2z(x, y) \\ &= \frac{\partial F}{\partial y} + \frac{\partial F}{\partial z} \frac{\partial z}{\partial y} \\ &= (3xz^2 - 10x^2yz) + (3z^2 + 6xyz - 5x^2y^2) \frac{\partial z}{\partial y}. \end{aligned}$$

Substituting x_0, y_0, z_0 and solving for the partial derivatives of z , we get

$$\begin{aligned} D_1z(x_0, y_0) &= -\frac{D_1F(x_0, y_0, z_0)}{D_3F(x_0, y_0, z_0)} = -\frac{32}{5}, \\ D_2z(x_0, y_0) &= -\frac{D_2F(x_0, y_0, z_0)}{D_3F(x_0, y_0, z_0)} = \frac{32}{5}. \end{aligned}$$

Thus, the tangent plane of the surface S at (x_0, y_0, z_0) is

$$32(x - 1) - 32(y + 1) + 5(z - 2) = 0.$$

Solving for z in terms of x and y gives the first-order Taylor expansion of $z = z(x, y)$ at (x_0, y_0) ; see Example 6.6.13.

Example 7.4.6. Consider the nonlinear system of polynomial equations

$$\begin{aligned}xu^2 + yzv + x^2z &= 3, \\xyv^3 + 2zu - u^2v^2 &= 2.\end{aligned}$$

We want to show that we can solve for u and v as smooth functions of x , y , and z in a neighborhood of the point $(1, 1, 1, 1, 1)$. Writing $\mathbf{x} = (x, y, z)$ and $\mathbf{y} = (u, v)$, we define the smooth map $F : \mathbb{R}^3 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$F(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} xu^2 + yzv + x^2z - 3 \\ xyv^3 + 2zu - u^2v^2 - 2 \end{bmatrix}.$$

Set $\mathbf{x}_0 = (1, 1, 1)$ and $\mathbf{y}_0 = (1, 1)$, and note that

$$D_2F(\mathbf{x}_0, \mathbf{y}_0) = \begin{bmatrix} 2xu & yz \\ 2z - 2uv^2 & 3xyv^2 - 2u^2v \end{bmatrix} \bigg|_{\mathbf{x}_0, \mathbf{y}_0} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix},$$

which is nonsingular (and thus has a bounded inverse). Therefore, by the implicit function theorem, we have that $\mathbf{y}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}))$ is a C^1 function in an open neighborhood of $\mathbf{x}_0$ satisfying $F(\mathbf{x}, \mathbf{y}(\mathbf{x})) = \mathbf{0}$.

Example 7.4.7.* For this example we introduce some notation. Let $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ be C^1 , where $i = 1, \dots, n$, and let $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. We write

$$\frac{\partial(f_1, f_2, \dots, f_n)}{\partial(x_1, x_2, \dots, x_n)} = \det \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}. \quad (7.24)$$

This is called the *Jacobian determinant* of the functions $f_1, f_2, \dots, f_n$.

Consider the system

$$\begin{aligned}f(x, y, z) &= 0, \\g(x, y, z) &= 0,\end{aligned}$$

where f and g are real-valued C^1 functions on $\mathbb{R}^3$ satisfying $J = \frac{\partial(f, g)}{\partial(y, z)} \neq 0$. By the implicit function theorem, we can solve for $y(x)$ and $z(x)$ as C^1 functions of x . Thus, we have $f(x, y(x), z(x)) = 0$ and $g(x, y(x), z(x)) = 0$. Taking the derivative yields

$$\begin{aligned}D_1f(x, y(x), z(x)) + D_2f(x, y(x), z(x))y'(x) + D_3f(x, y(x), z(x))z'(x) &= 0, \\D_1g(x, y(x), z(x)) + D_2g(x, y(x), z(x))y'(x) + D_3g(x, y(x), z(x))z'(x) &= 0.\end{aligned}$$

Moreover, from Cramer's rule (Corollary 2.9.24) we can solve for the derivatives $y'(x)$ and $z'(x)$ to get

$$y'(x) = J^{-1} \frac{\partial(f, g)}{\partial(z, x)} \quad \text{and} \quad z'(x) = J^{-1} \frac{\partial(f, g)}{\partial(x, y)};$$

see Exercise 7.25 for details.

7.4.2 Inverse Function Theorem

The implicit function theorem has another incarnation, called the *inverse function theorem*. It states that if the derivative of a function is invertible (or, rather, has a bounded inverse) at some point, then the function itself is invertible in a neighborhood of the image of that point. This should not come as a surprise—a differentiable function is very nearly equal to its derivative in a small neighborhood, so if the derivative is invertible in that neighborhood, the function should also be.

The inverse function theorem follows from the implicit function theorem, but the implicit function theorem can also be proved from the inverse function theorem. Depending on the application, one of these theorems may be easier to use than the other.

Theorem 7.4.8 (Inverse Function Theorem). *Assume that $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces, that U and V are open neighborhoods of $\mathbf{x}_0 \in X$ and $\mathbf{y}_0 \in Y$, respectively, and that $f : U \rightarrow V$ is a C^k map for some $k \in \mathbb{Z}^+$ satisfying $f(\mathbf{x}_0) = \mathbf{y}_0$. If $Df(\mathbf{x}_0) \in \mathcal{B}(X; Y)$ has a bounded inverse, then there exist open neighborhoods $U_0 \subset U$ of $\mathbf{x}_0$ and $V_0 \subset V$ of $\mathbf{y}_0$, and a unique C^k function $g : V_0 \rightarrow U_0$ that is inverse to f . In other words, $f(g(\mathbf{y})) = \mathbf{y}$ for all $\mathbf{y} \in V_0$ and $g(f(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x} \in U_0$. Moreover, for all $\mathbf{x} \in U_0$, we have*

$$Dg(\mathbf{y}) = Df(g(\mathbf{y}))^{-1}. \quad (7.25)$$

Proof. Define $F(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}$. Since $D_1 F(\mathbf{x}_0, \mathbf{y}_0) = Df(\mathbf{x}_0)$ has a bounded inverse, the implicit function theorem guarantees the existence of a neighborhood $U_1 \times V_0 \subset U \times X$ of the point $(\mathbf{x}_0, \mathbf{y}_0)$ and a C^k function $g : V_0 \rightarrow U_1$ such that $f(g(\mathbf{y})) = \mathbf{y}$ for all $\mathbf{y} \in V_0$, which implies that g is injective (see Theorem A.2.19). By restricting the codomain of g to $U_0 = g(V_0)$, we have that g is bijective. By Corollary A.2.20, this implies that $f : U_0 \rightarrow V_0$ and $g : V_0 \rightarrow U_0$ are inverses of each other. Note that $U_0 = U_1 \cap f^{-1}(V_0)$, which implies that U_0 is open. Finally, (7.25) follows by differentiating $f(g(\mathbf{y})) = \mathbf{y}$. $\square$

Example 7.4.9. The function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(t) = \cos(t)$ has $Df(t) = -\sin(t)$, which is nonzero whenever $t \neq k\pi$ for all $k \in \mathbb{Z}$. The inverse function theorem guarantees that for any point $t \neq k\pi$ there is a neighborhood $U_0 \subset \mathbb{R}$ of t , a neighborhood $V_0 \subset \mathbb{R}$ of $\cos(t)$, and an inverse function $g : V_0 \rightarrow U_0$. There cannot be a global inverse function $g : \mathbb{R} \rightarrow \mathbb{R}$ because f is not injective,

and the image of f lies in $[-1, 1]$; so the inverse function can only be defined on neighborhoods in $(-1, 1)$.

As an example, for $t \in (0, \pi)$ we can take neighborhoods $U_0 = (0, \pi)$ and $V_0 = (-1, 1)$ and let $g(x) = \text{Arccos}(x)$. If $x = \cos(t)$, then the inverse function theorem guarantees that the derivative of g is

$$Dg(x) = Df(t)^{-1} = \frac{1}{-\sin(t)} = \frac{-1}{\sqrt{1 - \cos^2(t)}} = \frac{-1}{\sqrt{1 - x^2}}.$$

Nota Bene 7.4.10. Whenever you use the inverse function theorem, there are likely to be a lot of negative exponents flying around. Some of these denote function inverses, and some of them denote matrix inverses, including the reciprocals of a scalars (which can be thought of as 1×1 matrices).

If the inverse function is $g = f^{-1}$, then the derivative $Dg(\mathbf{y}) = D(f^{-1})(\mathbf{y})$ of the inverse function is the matrix inverse of the derivative:

$$D(f^{-1})(\mathbf{y}) = (Df(\mathbf{x}))^{-1},$$

where f^{-1} on the left means the inverse function g , but the exponent on the right means the inverse of the matrix $Df(\mathbf{x})$.

Of course the inverse of a matrix is the matrix representing the inverse of the corresponding linear operator, so these exponents really are denoting the inverse function in both cases—the real problem is that many people confuse the linear operator $Df(\mathbf{x}) : X \rightarrow Y$ (the function we want to take the inverse of) with the nonlinear function $Df : U \rightarrow \mathcal{B}(X, Y)$, which often has no inverse at all.

Example 7.4.11. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the coordinate change in the plane from polar to Cartesian coordinates, given by $(r, \theta) \mapsto (r \cos \theta, r \sin \theta)$. Since

$$\det Df(r, \theta) = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r,$$

the function f is invertible in a neighborhood of (r, θ) whenever $r \neq 0$.

Note that no one function is the inverse of f on the entire punctured plane $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. After all, the function f is not injective, so it can't have a global inverse. But for points in the open set $U_0 = \{(r, \theta) \mid r > 0, \theta \in (-\pi/2, \pi/2)\}$, the image of f is the right half plane $V_0 = \{(x, y) \mid x > 0\}$, and on V_0 , the inverse function g is given by $(x, y) \mapsto (\sqrt{x^2 + y^2}, \text{Arctan}(y/x))$.

To find the derivative of the inverse function, one could try to find the inverse function explicitly in a neighborhood and then differentiate it. But the inverse function theorem tells us the derivative of the inverse without actually finding the inverse. In this example, if g is the inverse function, and

if $(x, y) = (r \cos \theta, r \sin \theta)$, then we have

$$Dg(x, y) = Df^{-1}(r, \theta) = \frac{1}{r} \begin{bmatrix} r \cos \theta & r \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ -\frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix}.$$

You can verify that this agrees with the result of finding g explicitly and then differentiating.

Theorem 7.4.12. *The inverse and implicit function theorems are equivalent.*

Proof. Since we used the implicit function theorem to prove the inverse function theorem, it suffices to prove the implicit function theorem, using the inverse function theorem. Let $G : U \times V \rightarrow X \times Z$ be given by $G(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, F(\mathbf{x}, \mathbf{y}))$. Note that

$$DG(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} I & 0 \\ D_1F(\mathbf{x}, \mathbf{y}) & D_2F(\mathbf{x}, \mathbf{y}) \end{bmatrix},$$

which has a bounded inverse whenever $D_2F(\mathbf{x}, \mathbf{y})$ has a bounded inverse. Applying the inverse function theorem to the equation $G(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{0})$, we have the solution $(\mathbf{x}, \mathbf{y}(\mathbf{x})) = G^{-1}(\mathbf{x}, \mathbf{0})$, which satisfies $F(\mathbf{x}, \mathbf{y}(\mathbf{x})) = \mathbf{0}$. $\square$

7.4.3 * Application: Navigation

Global positioning systems (GPS) are found in cars, phones, and many other devices. The implicit function theorem answers an important question about the design of a GPS, namely, how accurately time must be kept in order to achieve the desired accuracy in location.

To determine location, the GPS device measures the distance from the device to each of four satellites. If (x_i, y_i, z_i) is the location of the satellite S_i , and if r_i is its distance to the GPS device, then the location of the GPS device lies at the intersection of the following four spheres:

$$\begin{aligned} (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 &= r_1^2, \\ (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 &= r_2^2, \\ (x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2 &= r_3^2, \\ (x - x_4)^2 + (y - y_4)^2 + (z - z_4)^2 &= r_4^2. \end{aligned}$$

The distance r_i is calculated by considering the difference between the satellite time and the user time, or in other words $r_i = c(\Delta t_i - \Delta t_{i,\text{prop}})$, where $\Delta t_{i,\text{prop}}$ is the atmospheric propagation delay and c is the speed of light.

Unfortunately, Δt_i is difficult to compute, since the clocks on both the satellite and the GPS device tend to drift. The value of Δt_i is approximated as

$$\Delta t_i = \Delta t_{i,\text{fake}} + \Delta t_{i,\text{sat}} + \Delta t_{\text{loc}},$$

where $\Delta t_{i,\text{fake}}$ is the nominal time difference between the satellite clock and the local GPS clock, $\Delta t_{i,\text{sat}}$ is the drift in the satellite clock, and Δt_{loc} is the drift in

the local GPS unit clock. Hence, r_i can be written as

$$c(\Delta t_{i,\text{fake}} + \Delta t_{i,\text{sat}} - \Delta t_{i,\text{prop}} + \Delta t_{\text{loc}}).$$

To make things cleaner we write

$$t_i = \Delta t_{i,\text{fake}} + \Delta t_{i,\text{sat}} - \Delta t_{i,\text{prop}}$$

and

$$\ell = \Delta t_{\text{loc}},$$

and we let $\mathbf{t} = [t_1 \ \dots \ t_4]^\top$ and $\mathbf{x} = [x \ y \ z \ \ell]^\top$. Finally, let

$$F(\mathbf{t}, \mathbf{x}) = \begin{bmatrix} (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 - c^2(t_1 + \ell)^2 \\ (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 - c^2(t_2 + \ell)^2 \\ (x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2 - c^2(t_3 + \ell)^2 \\ (x - x_4)^2 + (y - y_4)^2 + (z - z_4)^2 - c^2(t_4 + \ell)^2 \end{bmatrix},$$

so the system of equations becomes

$$F(\mathbf{t}, \mathbf{x}) = \mathbf{0}.$$

We treat this as a system of four equations with four unknowns (x , y , z , and ℓ).

Suppose we wish to determine the change in $\mathbf{x}$ if we perturb $\mathbf{t}$, or, conversely, suppose we want $\mathbf{x}$ to be determined with a certain degree of precision. How much error can $\mathbf{t}$ have? The implicit function theorem is perfectly suited to give this kind of information. It states that $\mathbf{x}$ is a function of $\mathbf{t}$ if $D_2 F(\mathbf{t}, \mathbf{x})$ is invertible, and in that case we must have $D\mathbf{x}(\mathbf{t}) = -D_2 F(\mathbf{t}, \mathbf{x}(\mathbf{t}))^{-1} D_1 F(\mathbf{t}, \mathbf{x}(\mathbf{t}))$. Written more explicitly, we have

$$\begin{bmatrix} \frac{\partial x}{\partial t_1} & \frac{\partial x}{\partial t_2} & \frac{\partial x}{\partial t_3} & \frac{\partial x}{\partial t_4} \\ \frac{\partial y}{\partial t_1} & \frac{\partial y}{\partial t_2} & \frac{\partial y}{\partial t_3} & \frac{\partial y}{\partial t_4} \\ \frac{\partial z}{\partial t_1} & \frac{\partial z}{\partial t_2} & \frac{\partial z}{\partial t_3} & \frac{\partial z}{\partial t_4} \\ \frac{\partial \ell}{\partial t_1} & \frac{\partial \ell}{\partial t_2} & \frac{\partial \ell}{\partial t_3} & \frac{\partial \ell}{\partial t_4} \end{bmatrix} = - \begin{bmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} & \frac{\partial F_1}{\partial z} & \frac{\partial F_1}{\partial \ell} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} & \frac{\partial F_2}{\partial z} & \frac{\partial F_2}{\partial \ell} \\ \frac{\partial F_3}{\partial x} & \frac{\partial F_3}{\partial y} & \frac{\partial F_3}{\partial z} & \frac{\partial F_3}{\partial \ell} \\ \frac{\partial F_4}{\partial x} & \frac{\partial F_4}{\partial y} & \frac{\partial F_4}{\partial z} & \frac{\partial F_4}{\partial \ell} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial F_1}{\partial t_1} & \frac{\partial F_1}{\partial t_2} & \frac{\partial F_1}{\partial t_3} & \frac{\partial F_1}{\partial t_4} \\ \frac{\partial F_2}{\partial t_1} & \frac{\partial F_2}{\partial t_2} & \frac{\partial F_2}{\partial t_3} & \frac{\partial F_2}{\partial t_4} \\ \frac{\partial F_3}{\partial t_1} & \frac{\partial F_3}{\partial t_2} & \frac{\partial F_3}{\partial t_3} & \frac{\partial F_3}{\partial t_4} \\ \frac{\partial F_4}{\partial t_1} & \frac{\partial F_4}{\partial t_2} & \frac{\partial F_4}{\partial t_3} & \frac{\partial F_4}{\partial t_4} \end{bmatrix}.$$

Elementary calculations show that if the times are perturbed, the change in x is approximately

$$\Delta x \approx \frac{\partial x}{\partial t_1} \Delta t_1 + \frac{\partial x}{\partial t_2} \Delta t_2 + \frac{\partial x}{\partial t_3} \Delta t_3 + \frac{\partial x}{\partial t_4} \Delta t_4.$$

Moreover, one can show that if all of the Δt_i values are correct to within a nano-second (typical for the clocks used on satellites), then the coordinates will be correct to within 3 meters. For further information, see [NJN98].

7.5 Conditioning

If the answer is highly sensitive to perturbations, you have probably asked the wrong question.

—Nick Trefethen

Nearly every problem in applied mathematics can ultimately be expressed as a function. Solving these problems amounts to evaluating the functions. But when evaluating functions numerically, there are several potential sources of error. Two of the most important of these are errors in the inputs and errors in the intermediate computations. Since every measurement is inherently imprecise, and most numbers cannot be represented exactly as a floating-point number, inputs almost always have minor errors. Similarly, floating-point arithmetic almost always introduces minor errors at each intermediate computation, and depending on the algorithm, these can accumulate to produce significant errors in the output.

For each problem we must ask how much error can accumulate from round-off (floating-point) error in the algorithm, and how sensitive the function is to small changes in the inputs. The answer to the first question is measured by the *stability* of the algorithm. Stability is treated in Volume 2. The answer to the second question is captured in the *conditioning* of the problem. If a small change to the input only results in a small change to the output of the function, we say that the function is *well conditioned*. But if small changes to the input result in large changes to the output, we say the function is *ill conditioned*. Not surprisingly, a function can be ill conditioned for some inputs and well conditioned for other inputs.

Example 7.5.1. Consider the function $y = x/(1 - x)$. For values of x close to 1, a small change in x produces a large change in y . For example, if the correct input is $x = 1.001$, and if that is approximated by $\tilde{x} = 1.002$, the actual output of $\tilde{y} = 1/(1 - 1.002) = -500$ is very different from the desired output of $y = 1/(1 - 1.001) = -1000$. So this problem is ill conditioned near $x = 1$. Note that this error has nothing to do with round-off errors in the algorithm for computing the values—it is entirely a property of the problem itself.

But if the desired input is $x = 35$, then the correct output is $y = -1.0294$, and even a bad approximation to the input like $\tilde{x} = 36$ gives a good approximate output $\tilde{y} = -1.0286$. So this problem is well conditioned near 35.

7.5.1 Condition Number of a Function

The condition number of a problem measures how sensitive the problem is to changes in input values. In essence we would like something like

$$\text{change in output} = \text{condition number} \times \text{change in input}.$$

As shown in Example 7.5.1, such a number also depends on the input.

Definition 7.5.2. Let X and Y be normed linear spaces, and let $f : X \rightarrow Y$ be a function. The absolute condition number of f at $\mathbf{x} \in X$ is

$$\hat{\kappa}(\mathbf{x}) = \lim_{\delta \rightarrow 0^+} \sup_{\|\mathbf{h}\| < \delta} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})\|}{\|\mathbf{h}\|}.$$

Proposition 7.5.3. *Let X and Y be Banach spaces, and let $U \subset X$ be an open set containing $\mathbf{x}$. If $f : U \rightarrow Y$ is differentiable at $\mathbf{x}$, then*

$$\hat{\kappa}(\mathbf{x}) = \|Df(\mathbf{x})\|. \quad (7.26)$$

Proof. By Lemma 6.4.1, for every $\varepsilon > 0$, if $\|\mathbf{h}\|$ is sufficiently small, then

$$\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\| \leq \varepsilon\|\mathbf{h}\|.$$

Using the triangle inequality and dividing by $\|\mathbf{h}\|$ gives

$$\left| \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})\|}{\|\mathbf{h}\|} - \|Df(\mathbf{x})\| \right| \leq \varepsilon.$$

Therefore,

$$\left| \lim_{\delta \rightarrow 0^+} \sup_{\|\mathbf{h}\| < \delta} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})\|}{\|\mathbf{h}\|} - \|Df(\mathbf{x})\| \right| \leq \varepsilon$$

for every $\varepsilon > 0$, and so (7.26) follows. $\square$

In most settings, relative error is more useful than absolute error. An error of 1 is tiny if the true answer is 10^{20} , but it is huge if the true answer is 10^{-20} . Relative error accounts for this difference. Since the condition number is really about the size of errors in the output, the relative condition number is usually a better measure of conditioning than the absolute condition number.

Definition 7.5.4. *Let X and Y be normed linear spaces, and let $f : X \rightarrow Y$ be a function. The relative condition number of f at $\mathbf{x} \in X$ is*

$$\kappa(\mathbf{x}) = \lim_{\delta \rightarrow 0^+} \sup_{\|\mathbf{h}\| < \delta} \left(\frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})\|}{\|f(\mathbf{x})\|} \bigg/ \frac{\|\mathbf{h}\|}{\|\mathbf{x}\|} \right) = \frac{\hat{\kappa}(\mathbf{x})}{\|f(\mathbf{x})\|/\|\mathbf{x}\|}. \quad (7.27)$$

Remark 7.5.5. A problem is *well conditioned* at $\mathbf{x}$ if the relative condition number is small. What we mean by “small” depends on the problem, of course. Similarly, the problem is *ill conditioned* if the relative condition number is large. Again, what is meant by “large” depends on the problem.

Nota Bene 7.5.6. Roughly speaking, we have

$$\text{rel. change in output} = \text{rel. condition number} \times \text{rel. change in input}.$$

This leads to a general rule of thumb that, *without any error in the algorithm itself*, we should expect to lose k digits of accuracy if the relative condition number is 10^k .

If f is differentiable, then Proposition 7.5.3 gives a formula for the relative condition number in terms of the derivative.

Corollary 7.5.7. *If X and Y are Banach spaces, if $U \subset X$ is an open set containing $\mathbf{x}$, and if $f : U \rightarrow Y$ is differentiable at $\mathbf{x}$, then*

$$\kappa(\mathbf{x}) = \frac{\|Df(\mathbf{x})\|}{\|f(\mathbf{x})\|/\|\mathbf{x}\|}. \quad (7.28)$$

Example 7.5.8.

- (i) Consider the function $f(x) = \frac{x}{1-x}$ of Example 7.5.1. For this function $Df(x) = (1-x)^{-2}$, and hence by (7.28) we have

$$\kappa = \frac{\|Df(\mathbf{x})\|}{\|f(\mathbf{x})\|/\|\mathbf{x}\|} = \frac{\left| \frac{1}{(1-x)^2} \right|}{\left| \frac{x}{1-x} \right|/|x|} = \left| \frac{1}{1-x} \right|.$$

This problem is well conditioned when x is far from 1, and poorly conditioned when $|1-x|$ is small.

- (ii) Given y , consider the problem of finding x on the curve $x^3 - x = y^2$. Setting $F(x, y) = x^3 - x - y^2$, we can rewrite this as the problem of finding x to satisfy $F(x, y) = 0$. Note that $D_x F = 3x^2 - 1$, so, provided that $x \neq \pm\sqrt{1/3}$, the implicit function theorem applies and guarantees that there is (locally) a function $x(y)$ such that $F(x(y), y) = 0$. Moreover, $Dx(y) = dx/dy = 2y/(3x^2 - 1)$. Therefore, near a point (x, y) on the curve with $x \neq \pm\sqrt{1/3}$, the relative condition number of this function is

$$\kappa = \frac{|2y/(3x^2 - 1)|}{|x|/|y|} = \frac{2y^2}{|3x^3 - x|} = \frac{2|x^3 - x|}{|x||3x^2 - 1|} = \frac{2|x^2 - 1|}{|3x^2 - 1|}.$$

This problem is ill conditioned when x is close to $\pm\sqrt{1/3}$ and well conditioned elsewhere.

7.5.2 Condition of Finding a Simple Root of a Polynomial

We now show, using the implicit function theorem, that varying the coefficients of a single-variable polynomial p causes the simple roots (those of multiplicity 1) of p to vary as a continuous function of the coefficients, and we calculate the condition number of that function.

Proposition 7.5.9. *Define $P : \mathbb{F}^{n+1} \times \mathbb{F} \rightarrow \mathbb{F}$ by $P(\mathbf{a}, x) = \sum_{i=0}^n a_i x^i$. For any given $\mathbf{b} \in \mathbb{F}^{n+1}$ and any simple root z of the polynomial $p(x) = P(\mathbf{b}, x)$, there is a neighborhood U of $\mathbf{b}$ in $\mathbb{F}^{n+1}$ and a continuously differentiable function $r : U \rightarrow \mathbb{F}$ with $r(\mathbf{b}) = z$ such that $P(\mathbf{a}, r(\mathbf{a})) = 0$ for all $\mathbf{a} \in U$. Moreover, the relative*

condition number of r as a function of the i th coefficient a_i , at the point $(\mathbf{b}, z)$, is

$$\kappa = \left| \frac{z^{i-1}b_i}{p'(z)} \right|. \quad (7.29)$$

Proof. A root z of a polynomial p is simple if and only if $p'(z) \neq 0$. Differentiating P at $(\mathbf{b}, z)$ with respect to x gives $D_x P(\mathbf{b}, z) = \sum_{i=1}^n i b_i z^{i-1} = p'(z)$. Because $p'(z)$ is invertible, the implicit function theorem guarantees the existence of a neighborhood U of $\mathbf{a}$ and a unique continuous function $r : U \rightarrow \mathbb{F}$ such that $r(\mathbf{b}) = z$ and such that $P(\mathbf{a}, r(\mathbf{a})) = 0$ for all $\mathbf{a} \in U$. Moreover, we have

$$Dr(\mathbf{b}) = -D_x P(\mathbf{b}, z)^{-1} D_{\mathbf{a}} P(\mathbf{b}, z) = -\frac{1}{p'(z)} [1 \quad z \quad z^2 \quad \dots \quad z^n].$$

Combining this with (7.28) shows that the relative condition number of r as a function of the i th coefficient a_i is given by (7.29). $\square$

Example 7.5.10. If the derivative $p'(z)$ is small, relative to the coefficient b_i , then the root-finding problem is ill conditioned. A classic example of this is the Wilkinson polynomial

$$w(x) = \prod_{r=1}^{20} (x - r) = x^{20} - 210x^{19} + 20615x^{18} - \dots.$$

Perturbing the polynomial by changing the x^{19} -coefficient from -210 to -210.0000001 changes the roots substantially, as shown in Figure 7.3. This is because the derivative $p'(z)$ is small for roots like $z = 15$, relative to $z^{18}b_{19}$, where b_{19} is the coefficient of x^{19} . For example, at $z = 15$ we have

$$\kappa = \frac{15^{18}(-210)}{p'(15)} \approx 3.0 \times 10^{10}.$$

7.5.3 Condition Number of a Matrix

In this section we compute the condition number for problems of the form $A\mathbf{x} = \mathbf{b}$. There are several cases to consider:

- (i) Given $A \in M_n(\mathbb{F})$, what is the relative condition number of $f(\mathbf{x}) = A\mathbf{x}$?
- (ii) Given $\mathbf{x} \in \mathbb{F}^n$, what is the relative condition number of $g(A) = A\mathbf{x}$?
- (iii) Given $A \in M_n(\mathbb{F})$, what is the relative condition number of $h(\mathbf{b}) = A^{-1}\mathbf{b}$?

Although the relative condition numbers of these three cases are not identical, they are all bounded by the number $\|A\|\|A^{-1}\|$, and this is the best uniform bound, as we show below.

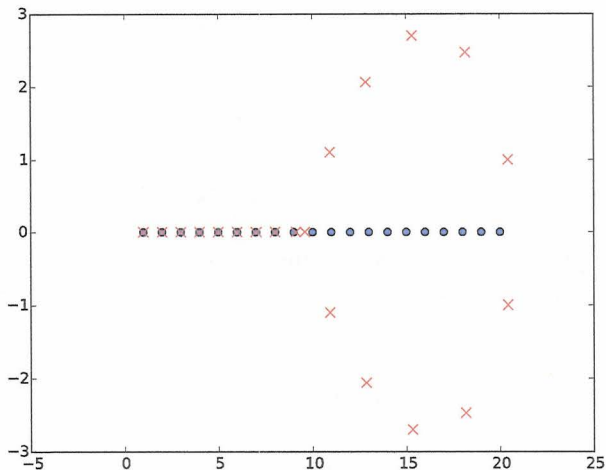


Figure 7.3. The blue dots are the roots of the Wilkinson polynomial $w(x)$ plotted in the complex plane. The red crosses are the roots of the polynomial perturbed by 10^{-7} in the x^{19} -coefficient. As described in Example 7.5.10, the roots are very sensitive to tiny variations in this coefficient because the relative condition number is very large.

Theorem 7.5.11.

- (i) If $A \in M_n(\mathbb{F})$ is nonsingular, the relative condition number of $f(\mathbf{x}) = A\mathbf{x}$ satisfies

$$\kappa = \|A\| \frac{\|\mathbf{x}\|}{\|A\mathbf{x}\|} \leq \|A\| \|A^{-1}\|. \quad (7.30)$$

Moreover, if the norm $\|\cdot\|$ is the 2-norm, then equality holds when $\mathbf{x}$ is a right singular vector of A corresponding to the minimal singular value.

- (ii) Given $\mathbf{x} \in \mathbb{F}^n$, the relative condition number of $g(A) = A\mathbf{x}$ satisfies

$$\kappa = \|\mathbf{x}\| \frac{\|A\|}{\|A\mathbf{x}\|} \leq \|A\| \|A^{-1}\| \quad (7.31)$$

if A is nonsingular. Moreover, for the 2-norm, equality holds when $\mathbf{x}$ is a right singular vector of A corresponding to the minimal singular value.

- (iii) If $A \in M_n(\mathbb{F})$ is nonsingular, the relative condition number of $h(\mathbf{b}) = A^{-1}\mathbf{b}$ satisfies

$$\kappa = \|A^{-1}\| \frac{\|\mathbf{b}\|}{\|A^{-1}\mathbf{b}\|} \leq \|A\| \|A^{-1}\|. \quad (7.32)$$

Moreover, for the 2-norm, equality holds when $\mathbf{b}$ is a left singular vector of A corresponding to the maximal singular value.

Proof. For (i) use (7.28) to find

$$\kappa = \frac{\|Df(\mathbf{x})\|}{\|A\mathbf{x}\|/\|\mathbf{x}\|} = \frac{\|A\|\|\mathbf{x}\|}{\|A\mathbf{x}\|}.$$

To get the upper bound, substitute $\mathbf{x} = A\mathbf{y}$ into the definition of $\|A^{-1}\|$ to get

$$\|A^{-1}\| = \sup_{\mathbf{x}} \frac{\|A^{-1}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{y}} \frac{\|A^{-1}A\mathbf{y}\|}{\|A\mathbf{y}\|} = \sup_{\mathbf{y}} \frac{\|\mathbf{y}\|}{\|A\mathbf{y}\|}.$$

This gives $\|A^{-1}\| \geq \|\mathbf{y}\|/\|A\mathbf{y}\|$ for all $\mathbf{y}$, from which we get (7.30).

See Exercise 7.27 for the proof that equality occurs when $\mathbf{x}$ is a right singular vector associated to the minimal singular value.

For (ii) it is straightforward to verify that for any $H \in M_n(\mathbb{F})$ we have that $Dg(A)H = H\mathbf{x}$. Exercise 3.29 gives $\|Dg(A)\| = \|\mathbf{x}\|$, so (7.28) gives

$$\kappa = \frac{\|\mathbf{x}\|}{\|g(A)\|/\|A\|} = \frac{\|\mathbf{x}\|\|A\|}{\|A\mathbf{x}\|} \leq \|A\|\|A^{-1}\|.$$

The same argument as for (i) gives equality for a singular vector associated to the minimal singular value.

Finally, substituting A^{-1} for A in (i) gives (iii), where equality holds for a right singular vector of $A^{-1} = V\Sigma^{-1}U^H$ associated to the minimal singular value $1/\sigma_1$ of A^{-1} , which is a left singular vector of A associated to maximal singular value σ_1 . $\square$

The previous theorem inspires the following definition.

Definition 7.5.12. Let $A \in M_n(\mathbb{F})$. The condition number of A is

$$\kappa(A) = \|A\|\|A^{-1}\|.$$

Nota Bene 7.5.13. Although $\kappa(A)$ is called the *condition number of the matrix* A , it is not the condition number (as given in Definition 7.5.4) of most problems associated to A . Rather, it is the supremum of the condition numbers of each of the various problems in Theorem 7.5.11; in other words, it is a sharp uniform bound for each of those condition numbers. Also the problem of finding eigenvalues of A has an entirely different condition number (see Section 7.5.4).

7.5.4 * Condition Number of Finding Simple Eigenvalues

Given a matrix $A \in M_n(\mathbb{F})$ and any simple eigenvalue of A , we show, using the implicit function theorem, that a continuous deformation of the entries of A continuously deforms the eigenvalue. We are also interested in the condition number of the problem of finding the eigenvalue.

Proposition 7.5.14. Let λ be a simple eigenvalue (that is, with algebraic multiplicity 1) of A with right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}^H$, both of norm 1. Fix

$E \in M_n(\mathbb{F})$ with $\|E\| = 1$. There is a neighborhood U of $0 \in \mathbb{R}$ and continuously differentiable functions $\xi(t), \mu(t)$ defined on U with $\xi(0) = \mathbf{x}$ and $\mu(0) = \lambda$ such that $\mu(t)$ is an eigenvalue of $A + tE$ with eigenvector $\xi(t)$.

Moreover, the absolute condition number of $\mu(t)$ at $t = 0$ satisfies

$$\hat{\kappa} = \frac{\mathbf{y}^H E \mathbf{x}}{\mathbf{y}^H \mathbf{x}} \leq \frac{1}{\mathbf{y}^H \mathbf{x}}, \quad (7.33)$$

and equality holds if $E = \mathbf{y} \mathbf{x}^H$.

Proof. Let $F : \mathbb{F} \times \mathbb{F}^n \times \mathbb{C} \rightarrow \mathbb{F}^n \times \mathbb{R}$ be given by

$$F(t, \xi, \mu) = \begin{bmatrix} (A + tE)\xi - \mu\xi \\ \xi^H \xi - 1 \end{bmatrix}.$$

We have $F(t, \xi, \mu) = [\mathbf{0} \ 0]^T$ if and only if $\|\xi\| = 1$ and ξ is an eigenvector of $A + tE$ with eigenvalue μ . In particular, we have $F(0, \mathbf{x}, \lambda) = 0$. Computing the derivative with respect to the last two coordinates ξ, μ , we have

$$D_{\xi, \mu} F = \begin{bmatrix} A + tE - \mu I & -\xi \\ 2\xi^H & 0 \end{bmatrix}.$$

Evaluating at $(0, \mathbf{x}, \lambda)$ gives

$$D_{\xi, \mu} F(0, \mathbf{x}, \lambda) = \begin{bmatrix} A - \lambda I & -\mathbf{x} \\ 2\mathbf{x}^H & 0 \end{bmatrix}.$$

Exercise 7.31 shows that $D_{\xi, \mu} F(0, \mathbf{x}, \lambda)$ is invertible. The implicit function theorem now applies to guarantee the existence of a differentiable function $f(t) = (\xi(t), \mu(t)) \in \mathbb{F}^n \times \mathbb{C}$, defined in a neighborhood U of $t = 0$ such that

$$F(t, \xi(t), \mu(t)) = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix} \quad (7.34)$$

for all $t \in U$, and thus $\mu(t)$ is an eigenvalue of $A + tE$ with eigenvector $\xi(t)$.

Differentiating (7.34) with respect to t at $t = 0$ gives

$$\begin{bmatrix} (E - \mu'(0))\mathbf{x} + (A - \lambda I)\xi'(0) \\ 2\mathbf{x}^H \xi'(0) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}.$$

Multiplication of the top row on the left by $\mathbf{y}^H$ gives

$$\mathbf{0} = \mathbf{y}^H (E - \mu'(0))\mathbf{x} + \mathbf{y}^H (A - \lambda I)\xi'(0) = \mathbf{y}^H E \mathbf{x} - \mu'(0)\mathbf{y}^H \mathbf{x},$$

which gives

$$\hat{\kappa} = \mu'(0) = \frac{\mathbf{y}^H E \mathbf{x}}{\mathbf{y}^H \mathbf{x}} \leq \frac{1}{\mathbf{y}^H \mathbf{x}},$$

where the last inequality follows from the fact that $\|E\| = 1$. In the special case that $E = \mathbf{y} \mathbf{x}^H$, it is immediate that the inequality is an equality. $\square$

As a corollary, we see that finding simple eigenvalues of normal matrices is especially well conditioned.

Corollary 7.5.15. *If A is normal, then the absolute condition number of finding any simple eigenvalue is no greater than 1.*

Proof. If A is normal, then the second spectral theorem (or rather the analogue of Corollary 4.4.9 for normal matrices) guarantees that there is an orthonormal eigenbasis of A . Since λ is simple, one of the basis elements $\mathbf{x}$ corresponds to λ , and by Remark 4.3.19 $\mathbf{x}^H$ is a corresponding left eigenvector. Thus, $\mathbf{y}^H \mathbf{x} = \mathbf{x}^H \mathbf{x} = 1$, and $\hat{\kappa} \leq 1$. $\square$

Example 7.5.16. When a matrix A is not normal, calculation of eigenvalues can be ill conditioned. For example, consider the matrix

$$A = \begin{bmatrix} 1 & 1000 \\ 0.001 & 1 \end{bmatrix},$$

which has eigenvalues $\{0, 2\}$. Take as right eigenvector $\mathbf{x} = [1 \quad -0.001]^T$ and left eigenvector $\mathbf{y}^H = [-0.001 \quad 1]$, and note that $\|\mathbf{x}\|_\infty = \|\mathbf{y}\|_\infty = 1$. Setting

$$E = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

we have by Proposition 7.5.14 that there is a continuous function $\mu(t)$ such that $\mu(0) = 0$ and $\mu(t)$ is an eigenvalue of $A + tE$ near $t = 0$. Moreover, (7.33) gives the absolute condition number $\hat{\kappa}$ for μ at $t = 0$ as

$$\hat{\kappa} = \frac{\mathbf{y}^H E \mathbf{x}}{\mathbf{y}^H \mathbf{x}} = \frac{1}{2 \times 10^{-3}} = 500.$$

The ill conditioning is illustrated by looking at the change in μ from $t = 0$ to $t = -0.001$. When $t = -0.001$, the matrix

$$A + tE = \begin{bmatrix} 1 & 1000 \\ 0 & 1 \end{bmatrix}$$

has a double eigenvalue $\mu(-0.001) = 1$; so a small deformation in t results in a relatively large change in μ .

Vista 7.5.17. In Chapter 14 we discuss *pseudospectra* of matrices. One of the many uses of pseudospectra is that they help us better understand and quantify some of the problems associated with ill conditioning of eigenvalues for nonnormal matrices.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 7.1. Consider the closed, complete metric space $X = [0, \infty)$ with the usual metric $d(x, y) = |x - y|$. Show that the map

$$f(x) = \frac{x + \sqrt{x^2 + 1}}{2}$$

satisfies $d(f(x), f(y)) < d(x, y)$ for all $x \neq y \in X$ and yet has no fixed point. Why doesn't this violate the contraction mapping principle?

- 7.2. Consider the sequence $(x_k)_{k=0}^\infty$ defined by the recursion $x_n = \sqrt{\alpha + x_{n-1}}$, with $\alpha > 1$ and $x_0 = 1$. Prove that

$$\lim_{n \rightarrow \infty} x_n = \frac{1 + \sqrt{1 + 4\alpha}}{2}.$$

Hint: Show that the function $f(x) = \sqrt{\alpha + x}$ is a contraction on the set $[0, \infty)$, with the usual metric.

- 7.3. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a continuous function satisfying $\|f(\mathbf{x})\| < k\|\mathbf{x}\|$ for some $k \in (0, 1)$ and for every $\mathbf{x} \neq \mathbf{0}$. For some initial point $\mathbf{x}_0 \in \mathbb{R}^n$ define the sequence $(\mathbf{x}_n)_{n=0}^\infty$ recursively by the rule $\mathbf{x}_{n+1} = f(\mathbf{x}_n)$. Show that $\mathbf{x}_n \rightarrow \mathbf{0}$. Hint: First prove that the sequence must have a limit, and show that if $\bar{\mathbf{x}}$ is the limit of the sequence, then $f(\bar{\mathbf{x}})$ is also. Use this to show that if $\bar{\mathbf{x}} \neq \mathbf{0}$, it cannot be the limit.
- 7.4. Let (X, d) be a metric space and $f : X \rightarrow X$ a contraction with constant K . Prove for $\mathbf{x}, \mathbf{y} \in X$ that

$$d(\mathbf{x}, \mathbf{y}) \leq \frac{1}{1 - K} (d(\mathbf{x}, f(\mathbf{x})) + d(\mathbf{y}, f(\mathbf{y}))).$$

Use this to prove that a contraction mapping can have at most one fixed point.

- 7.5. Let (X, d) be a metric space and $f : X \rightarrow X$ a contraction with constant $K < 1$. Using the inequality in the previous problem, prove that for any $\mathbf{x} \in X$ the sequence $(f^n(\mathbf{x}))_{n=0}^\infty$ satisfies the inequality

$$d(f^m(\mathbf{x}), f^n(\mathbf{x})) \leq \frac{K^m + K^n}{1 - K} d(\mathbf{x}, f(\mathbf{x})).$$

Use this to prove that $(f^n(\mathbf{x}))_{n=0}^\infty$ is a Cauchy sequence. Finally, prove that if $\bar{\mathbf{x}}$ is the limit of this sequence, then $f(\bar{\mathbf{x}}) = \bar{\mathbf{x}}$, and for any integer $n > 0$

$$d(f^n(\mathbf{x}), \bar{\mathbf{x}}) \leq \frac{K^n}{1 - K} d(\mathbf{x}, f(\mathbf{x})).$$

- 7.6. Prove Theorem 7.1.14. Hint: This follows the same ideas used to prove Theorem 7.1.8; define a sequence via the method of successive approximations and prove convergence.

- 7.7. Let $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(x, y) = \cos(\cos(x)) + y$. Show that $f(x, y)$ is a C^∞ uniform contraction mapping. Hint: Consider using the mean value theorem and the fact that $|\sin(x)| \leq \sin(1) < 1$ for all $x \in [-1, 1]$.
- 7.8. Let $C_b([0, \infty); \mathbb{R}) = \{g \in C([0, \infty); \mathbb{R}) : \|g\|_{L^\infty} < \infty\}$ be the set of continuous functions with bounded sup norm. Given $\mu > 0$ let $T_\mu : C_b([0, \infty); \mathbb{R}) \times C_b([0, \infty); \mathbb{R}) \rightarrow C([0, \infty); \mathbb{R})$ be given by

$$(x, f) \mapsto T_{\mu, f}[x](t) = f(t) + \frac{\mu}{2} \int_0^t e^{-\mu s} x(s) ds.$$

- (i) Prove that $(C_b([0, \infty); \mathbb{R}), \|\cdot\|_{L^\infty})$ is a Banach space. Hint: The proof of Theorem 5.6.8 may be useful.
 - (ii) Prove that the image of $T_{\mu, f}$ lies in $C_b([0, \infty); \mathbb{R})$ for every $f \in C_b([0, \infty); \mathbb{R})$.
 - (iii) Prove that $T_\mu : C_b([0, \infty); \mathbb{R}) \times C_b([0, \infty); \mathbb{R}) \rightarrow C_b([0, \infty); \mathbb{R})$ is a uniform contraction.
- 7.9. Let $(X, \|\cdot\|_X)$ be a Banach space and $A \in \mathcal{B}(X)$ an operator with $\|A\| < 1$, where $\|\cdot\|$ is the induced norm on $\mathcal{B}(X)$.
- (i) Show that $f : X \times X \rightarrow X$ defined by $f(\mathbf{x}, \mathbf{y}) = A\mathbf{x} + \mathbf{y}$ is a C^∞ uniform contraction mapping.
 - (ii) Find the function $g(\mathbf{y})$ that sends each point of $\mathbf{y}$ to the unique fixed point of the corresponding contraction $f(\cdot, \mathbf{y})$, and verify that $g : X \rightarrow X$ is C^∞ .
 - (iii) Verify that the map T_μ of the previous exercise is of the form $A\mathbf{x} + \mathbf{y}$ with $\|A\| < 1$.

- 7.10. For $\alpha \in \mathbb{R}$, define the weighted norm on $C([0, \infty); \mathbb{R})$ as

$$\|f\|_\alpha = \sup_{t \geq 0} \{e^{\alpha t} |f(t)|\}$$

and the corresponding function space

$$C_\alpha = \{f \in C([0, \infty); \mathbb{R}) : \|f\|_\alpha < \infty\}.$$

Essentially the same argument as in Exercise 7.8(i) shows that $(C_\alpha, \|\cdot\|_\alpha)$ is a Banach space.

- (i) Fix $z \in C_\alpha$, and define a map $T_z : C_\alpha \times C_\alpha \rightarrow C([0, \infty); \mathbb{R})$ by

$$(x, f) \mapsto T_{z,f}[x](t) = z(t) + f(t) \int_t^\infty e^{(\alpha-1)s} x(s) ds.$$

Prove that for each $f \in C_\alpha$ the map $T_{z,f}$ is an operator on C_α .

- (ii) Fix $k < 1$ and let $B = \{f \in C_\alpha : \|f\|_\alpha < k\}$. Prove that $T_z : C_\alpha \times B \rightarrow C_\alpha$ is a uniform contraction.

- 7.11. Suppose $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces, $U \subset X$ and $V \subset Y$ are open, and the function $f : \overline{U} \times V \rightarrow \overline{U}$ is a uniform contraction with constant $0 \leq \lambda < 1$. In addition, suppose there exists a C such that

$$\|f(\mathbf{x}, \mathbf{b}_1) - f(\mathbf{x}, \mathbf{b}_2)\|_X \leq C\|\mathbf{b}_2 - \mathbf{b}_1\|_Y$$

for all $\mathbf{b}_1, \mathbf{b}_2 \in V$ and all $\mathbf{x} \in \overline{U}$. Show that the function $g : V \rightarrow \overline{U}$ that sends each $\mathbf{y} \in V$ to the unique fixed point of the contraction is Lipschitz.

- 7.12.* Let D be a closed subset of a Banach space, and let $f : D \rightarrow D$ be continuous. Choose $\mathbf{x}_0 \in D$ and let $\mathbf{x}_{n+1} = f(\mathbf{x}_n)$ for every integer $n \in \mathbb{N}$. Prove that if the series $\sum_{n=0}^\infty \|\mathbf{x}_{n+1} - \mathbf{x}_n\|$ converges (in $\mathbb{R}$), then the sequence $(\mathbf{x}_n)_{n=0}^\infty$ converges in D to a fixed point of f .

- 7.13. (i) Using Newton's method, derive the square root formula (7.15) in Example 7.3.5.
 (ii) Derive the cube root formula (7.16) in Example 7.3.7. Then compute the long division in (7.17) by hand to get 5 decimals of accuracy.
- 7.14. The generic proof of convergence in Newton's method requires a sufficiently close initial guess x_0 . Prove that the square root solver in (7.15) converges as long as $x \neq 0$.
- 7.15. Although Newton's method is very fast in most cases, there are situations where it converges very slowly. Suppose that

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The function f can be shown to be C^∞ , and 0 is the only solution of $f(x) = 0$. Show that if $x_0 = 0.0001$, it takes more than one hundred million iterations of the Newton method to get below 0.00005. Prove, moreover, that the closer x_n is to 0, the slower the convergence. Why doesn't this violate Theorem 7.3.4?

- 7.16. Let $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be given as

$$F(x, y) = \begin{bmatrix} x - y^2 + 8 + \cos y \\ y - x^2 + 9 + 2 \cos x \end{bmatrix}.$$

Using the initial guess $\mathbf{x}_0 = (\pi, \pi)$, compute the first iteration of Newton's method by hand.

- 7.17. Determine whether the previous sequence converges by checking the Newton–Kantorovich bound given in (7.20). Hint: To compute the operator norms $\|Df(\mathbf{x}_0)^{-1}\|_2$ and $\|Df(\mathbf{x}) - Df(\mathbf{y})\|_2$ you may wish to use the results of Exercise 3.28.

- 7.18. Consider the points $(x, y) \in \mathbb{R}^2$ on the curve $\cos(xy) = 1 + \tan(y)$. Find conditions on x and y that guarantee x is locally a function of y , and find dx/dy .
- 7.19. Find the total Fréchet derivative of z as a function of x and y on the surface $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + xyz + 4x^5z^3 + 6z + y = 0\}$ at the origin.
- 7.20. Show that the equations

$$\begin{aligned}\sin(x+z) + \ln(yz^2) &= 0, \\ e^{x+z} + yz &= 0\end{aligned}$$

implicitly define C^1 functions $x(z)$ and $y(z)$ in a neighborhood of the point $(1, 1, -1)$.

- 7.21. Show that there exist C^1 functions $f(x, y)$ and $g(x, y)$ in a neighborhood of $(0, 1)$ such that $f(0, 1) = 1$, $g(0, 1) = -1$, and

$$\begin{aligned}f(x, y)^3 + xg(x, y) - y &= 0, \\ g(x, y)^3 + yf(x, y) - x &= 0.\end{aligned}$$

- 7.22. The principal inverse secant function $f(x) = \text{Arcsec}(x)$ has domain $(-\infty, -1) \cup (1, \infty)$ and range $(0, \pi/2) \cup (\pi/2, \pi)$. Using only the derivative of the secant, basic trigonometric properties, and the inverse function theorem, prove that the derivative of f is

$$\frac{df}{dx} = \frac{1}{|x|\sqrt{x^2 - 1}}.$$

- 7.23. Let $S : M_2(\mathbb{R}) \rightarrow M_2(\mathbb{R})$ be given by $S(A) = A^2$. Does S have a local inverse in a neighborhood of the identity matrix? Justify your answer.
- 7.24.* Denote the functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ and $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ in the standard bases as

$$\begin{aligned}f(\mathbf{x}) &= (f_1(x_1, x_2), f_2(x_1, x_2)), \\ g(\mathbf{x}) &= (g_1(x_1, x_2), g_2(x_1, x_2)),\end{aligned}$$

where $\mathbf{x} = (x_1, x_2)$. Prove: If f and g are C^1 and satisfy $f(g(\mathbf{y})) = \mathbf{y}$ for all $\mathbf{y}$, then

$$\begin{aligned}\frac{dg_1}{dy_1} &= J^{-1} \frac{\partial f_2}{\partial x_2} \quad \text{and} \quad \frac{dg_1}{dy_2} = -J^{-1} \frac{\partial f_1}{\partial x_2}, \\ \frac{dg_2}{dy_1} &= -J^{-1} \frac{\partial f_2}{\partial x_1} \quad \text{and} \quad \frac{dg_2}{dy_2} = J^{-1} \frac{\partial f_1}{\partial x_1},\end{aligned}$$

where $J = \frac{\partial(f_1, f_2)}{\partial(x_1, x_2)}$; see (7.24) in Example 7.4.7.

- 7.25.* Work out the details of Example 7.4.7, showing that Cramer's rule gives the formulas for the derivatives.

- 7.26. Find the relative condition number at each x_0 in the domain of the following functions:
- (i) e^x .
 - (ii) $\ln(x)$.
 - (iii) $\cos(x)$.
 - (iv) $\tan(x)$.
- 7.27. Finish the proof of Theorem 7.5.11 by showing that in the 2-norm, if $\mathbf{x}$ is a right singular vector of A associated to the minimal singular value, then equality holds in (7.30).
- 7.28. Given $(x, y) \in \mathbb{C}^2$, consider the problem of finding z such that $x^2 + y^2 - z^3 + z = 0$. Find all the points (x, y) for which z is locally a function of (x, y) . For a fixed value of y , find the relative condition number of z as a function of x . What is this relative condition number near the point $(x, y) = (0, 0)$?
- 7.29. Give an example of a matrix A with condition number $\kappa(A) > 1000$ (assuming the 2-norm). Give an example of a matrix B with condition number $\kappa(B) = 1$. Are there any matrices with condition number less than 1? If so, give an example. If not, prove it.
- 7.30. Proposition 7.5.9 gives sufficient conditions for the roots of a polynomial to be a continuous function of the coefficients.
- (i) Consider the roots of the polynomial $f(a, x) = x^2 + a$, where $a, x \in \mathbb{R}$. If $a = 0$, then $x_0 = 0$ is a root, but if $a > 0$, then f has no roots in $\mathbb{R}$. Why doesn't this contradict Proposition 7.5.9?
 - (ii) The quadratic formula gives an explicit formula for all the roots of a quadratic polynomial as a function of the coefficients. There is a similar but much more complicated formula for roots of cubic and quartic polynomials, but Abel's theorem guarantees that there is no general algebraic formula for the roots of a polynomial of degree 5 or greater. Why doesn't this contradict Proposition 7.5.9?
- 7.31.* Prove that the derivative $D_{\xi, \mu} F(0, \mathbf{x}, \lambda)$ of

$$F(t, \xi, \mu) = \begin{bmatrix} (A + tE)\xi - \mu\xi \\ \xi^H \xi - 1 \end{bmatrix},$$

as described in the proof of Proposition 7.5.14, is invertible at $t = 0$. Hint: Schur's lemma guarantees the existence of an orthonormal matrix Q such that $Q^H A Q$ is upper triangular (of course, with all the eigenvalues of A on the diagonal). Conjugate $D_{\xi, \mu} F(0, \mathbf{x}, \lambda)$ by

$$\begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix}$$

to get a matrix that you should be able to identify as nonsingular. Question: How does this rely on the fact that λ is a simple eigenvalue?

Notes

For more on the uniform contraction mapping principle and further generalizations, see [Chi06, Sect. 1.11].

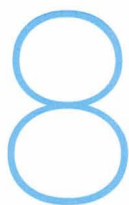
A readable description of the Newton–Kantorovich bound is given in [Ort68]. Kantorovich actually proved the result in two different ways, and an English translation of his two proofs is given in [Kan52, KA82].

Our treatment of conditioning is inspired by [TB97, Dem97, GVL13]. For more on conditioning in general, see [TB97, Dem97]. For more on conditioning of the eigenvalue problem, see [GVL13, Sect. 7.2.2].

Exercises 7.4–7.5 are based on the paper [Pal07]. Exercise 7.15 comes from [Ans06].

Part III

Nonlinear Analysis II



Integration I

I'm not for integration, and I'm not against it.
—Richard Pryor

Integration is an essential tool in applied mathematics. It is a fundamental operation in many problems arising in physics and mechanics, and it is also central to the modern treatment of probability.

Most treatments of modern integration begin with a long treatment of measure theory and then use that to develop the theory of integration. We take a completely different approach, based on some of the topological ideas developed earlier in this text. Our approach is unusual, but it has many advantages. We feel it is more natural and direct. For many students, this approach is more intuitive than the traditional treatments of Lebesgue integration. It also avoids most of the work involved in developing measure theory and instead leverages and reinforces many of the ideas developed in the first half of this text.

We begin our treatment of integration in this chapter by extending the definition of the regulated integral (Section 5.10) to multivariable functions taking values in a Banach space X . Recall that in Section 5.10 we defined the single-variable integral by defining the integral in the “obvious” way on step functions, and then using the continuous linear extension theorem (Theorem 5.7.6) to show that there is a unique linear extension of the integral operator from step functions to the closure (with respect to the L^∞ -norm) of the space of step functions. The construction of the multivariable regulated integral is essentially identical to the single-variable case. The hardest part is just defining and keeping track of notation in the definition of higher-dimensional step functions.

Just as in the single-variable case, continuous functions lie in the closure of the space of step functions, so this gives a definition of the integral, called the *regulated integral* for all continuous functions (on a compact interval). The regulated integral agrees with the Riemann integral whenever it is defined, but, unfortunately, both the Riemann and the regulated integrals have some serious drawbacks. The most important of these drawbacks is that these integrals do not behave well with respect to limits if the limits are not uniform. It is often desirable to move a limit past the integral sign (from $\lim_{n \rightarrow \infty} \int f_n$ to $\int \lim_{n \rightarrow \infty} f_n$) even when the limit is not uniform.

But unfortunately the function $\lim_{n \rightarrow \infty} f_n$ need not be Riemann integrable, even if the individual terms f_n are all Riemann integrable.

Instead of limits in the L^∞ -norm (uniform limits), we often need to consider limits in the L^1 -norm (given by $\|f\|_{L^1} = \int \|f\|$). The space of Riemann-integrable functions is not complete in the L^1 -norm, so there are L^1 -Cauchy sequences that do not converge in this space. To remedy this, we need to extend the space to a larger one that is complete with respect to the L^1 -norm, and then we can use the continuous linear extension theorem (Theorem 5.7.6) to extend the integral to this larger space. The result of all this is a much more general theory of integration.⁴²

If the functions being integrated take values in $X = \mathbb{R}$, then this construction is known as the *Daniell integral*. We will usually call it the *Lebesgue integral*, however, because it is equivalent to the Lebesgue integral, and that name is more familiar to most mathematicians. If X is a more general Banach space, then this construction is called (or, rather, is equivalent to) the *Bochner integral*. For simplicity, we restrict ourselves to the case of $X = \mathbb{R}$ for most of this chapter and the next, but much of what we do works just as well when X is a general Banach space.

It is important to keep in mind that on a bounded set all Riemann-integrable and regulated-integrable functions are also Lebesgue integrable, and for these functions the Riemann and regulated integrals are the same as the Lebesgue integral. Thus, to compute the Lebesgue integral of a continuous function on a compact set, for example, we can just use the usual techniques for finding the regulated integral of that function.

In this chapter we begin by extending the definition of the regulated integral to multivariable functions, and then by giving an overview of the main ideas and theorems of Lebesgue integration (in the next chapter we give the details and complete the proofs). The majority of the chapter is devoted to some of the main tools of integration. These include three important convergence theorems (the monotone convergence theorem, Fatou's lemma, and the dominated convergence theorem), Fubini's theorem, and a generalized change of variables formula. The three convergence theorems give useful conditions for when limits commute with integration. Fubini's theorem gives a way to convert multivariable integrals into several single-variable integrals (*iterated integrals*), which can then be evaluated in the usual ways. Fubini also shows us how to change the order in which those iterated single-variable integrals are evaluated, and in many situations changing the order of integration can greatly simplify the problem. Finally, the multivariable change of variables formula is analogous to the single-variable version (Corollary 6.5.6) and is also very useful.

8.1 Multivariable Integration

Recall that in Section 5.10 we define the single-variable integral by first defining integration in the obvious way for step functions, and then the continuous linear extension theorem (Theorem 5.7.6) gives a unique way to extend that definition to the closure of the space of step functions, with respect to the L^∞ -norm. The same construction works for multivariable integration, once we have a definition of n -dimensional intervals and step functions on $\mathbb{R}^n$.

⁴²The approach to integration in these next two chapters is inspired by, but is perhaps even more unusual than, what we did with integration in Chapter 5 (see Nota Bene 5.10.1).

8.1.1 Multivariable Step Functions

Step functions from a compact interval $[a, b] \subset \mathbb{R}$ to a Banach space $(X, \|\cdot\|)$ provide the key tool in the definition of the single-variable regulated integral. In this subsection we generalize these to multivariable step functions. To do this we must first define the idea of a multivariable interval.

Definition 8.1.1. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, with $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$. We denote by $[\mathbf{a}, \mathbf{b}]$ the closed n -interval (or box)

$$[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n.$$

To define step functions on $[\mathbf{a}, \mathbf{b}]$ we must define a generalized subdivision of the n -interval $[\mathbf{a}, \mathbf{b}]$. Recall from Definition 5.10.2 that a subdivision P of an interval $[a, b] \subset \mathbb{R}$ is a finite sequence $a = t^{(0)} < t^{(1)} < \cdots < t^{(k-1)} < t^{(k)} = b$.

Definition 8.1.2. A subdivision $\mathcal{P}$ of the n -interval $[\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^n$ consists of a subdivision $P_i = \{t_i^{(0)} = a_i < t_i^{(1)} < \cdots < t_i^{(k_i-1)} < t_i^{(k_i)} = b_i\}$ of the interval $[a_i, b_i]$ for each $i \in \{1, \dots, n\}$ (note that the lengths of the subdivisions in each dimension may vary); see Figure 8.1.

Definition 8.1.3. A subdivision of an n -interval $[\mathbf{a}, \mathbf{b}]$ gives a decomposition of $[\mathbf{a}, \mathbf{b}]$ into a union of partially open and closed subintervals as follows. Each $t_i^{(j)}$ with $0 < j < k_i$ defines a hyperplane $H_i^{(j)} = \{\mathbf{x} \in \mathbb{R}^n \mid x_i = t_i^{(j)}\}$ in $\mathbb{R}^n$, which divides the interval into two regions: a partially open interval

$$[a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [a_i, t_i^{(j)}) \times [a_{i+1}, b_{i+1}] \times \cdots \times [a_n, b_n]$$

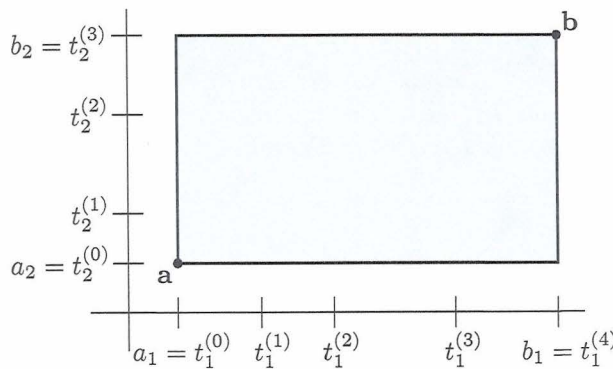


Figure 8.1. Partition of an interval $[\mathbf{a}, \mathbf{b}]$ in $\mathbb{R}^2$, consisting of two one-dimensional partitions: $a_1 = t_1^{(0)} < t_1^{(1)} < t_1^{(2)} < t_1^{(3)} < t_1^{(4)} = b_1$, and $a_2 = t_2^{(0)} < t_2^{(1)} < t_2^{(2)} < t_2^{(3)} = b_2$. See Definition 8.1.2 for details.

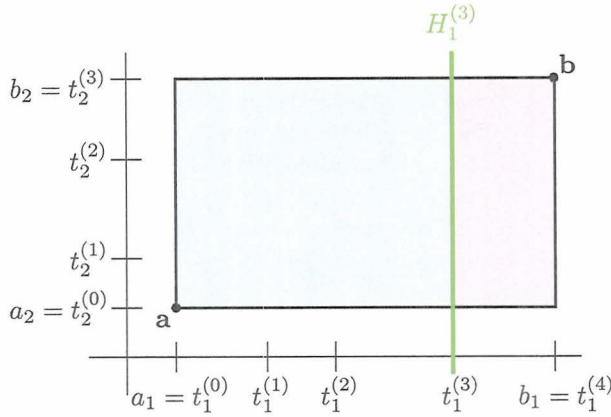


Figure 8.2. The hyperplane $H_1^{(3)}$ (green) in $\mathbb{R}^2$ defined by the point $t_1^{(3)}$ divides the interval into the union of a partially open interval (blue) and a closed interval (red), as described in Definition 8.1.3.

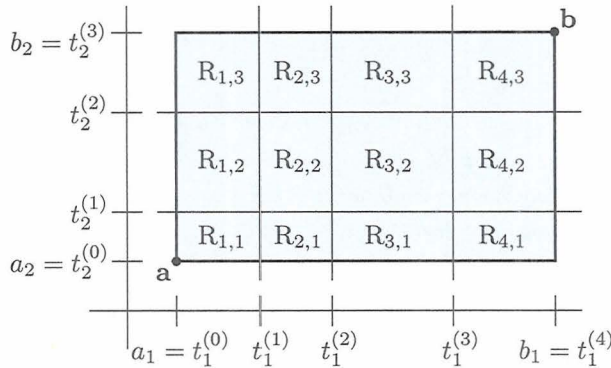


Figure 8.3. A partition of the interval $[a, b]$ in $\mathbb{R}^2$ divides $[a, b]$ into a union of subintervals R_I , some partially open and one closed (the upper right interval, containing b), as described in Definition 8.1.3.

and a closed interval

$$[a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [t_i^{(j)}, b_i] \times [a_{i+1}, b_{i+1}] \times \cdots \times [a_n, b_n]$$

(see Figure 8.2).

Repeating this process for each resulting interval (either partially open or closed) and for each hyperplane, we get a decomposition of $[a, b]$ into a union of intervals. Specifically, for each n -tuple $I = (i_1, \dots, i_n) \in \{1, \dots, k_1\} \times \cdots \times \{1, \dots, k_n\}$ of indices we get a subinterval R_I (see Figure 8.3).

The interval $[a, b]$ is the union

$$[a, b] = \bigcup_{I \in \mathcal{P}} R_I.$$

To simplify notation we write $I \in \mathcal{P}$ to denote that the index I lies in the product of the indices of $\mathcal{P}$, that is, $I \in \{1, \dots, k_1\} \times \dots \times \{1, \dots, k_n\}$. We use this notation repeatedly when discussing step functions.

Throughout the rest of this section, assume that $(X, \|\cdot\|)$ is a Banach space.

Definition 8.1.4. For any set $E \subset \mathbb{R}^n$ the indicator function $\mathbb{1}_E$ of E is the function

$$\mathbb{1}_E(z) = \begin{cases} 1, & z \in E, \\ 0, & z \notin E. \end{cases}$$

A function $s : [a, b] \rightarrow X$ is a step function on $[a, b]$ if there is a subdivision $\mathcal{P}$ of $[a, b]$ such that s can be written in the form

$$s(t) = \sum_{I \in \mathcal{P}} \mathbf{x}_I \mathbb{1}_{R_I}(t).$$

Here $\mathbf{x}_I \in X$ for each $I \in \mathcal{P}$, and the subintervals $R_I \subset [a, b]$ are determined by the index and the partition, as described in Definition 8.1.3. More generally, for any $E \subset \mathbb{R}^n$ we consider a function $s : E \rightarrow X$ to be a step function if it is zero outside of an interval $[a, b]$ and the restriction $s|_{[a, b]}$ to $[a, b]$ is a step function on $[a, b]$.

The proof of the next proposition is essentially identical to its one-dimensional counterpart (Proposition 5.10.3).

Proposition 8.1.5. The set $S([a, b]; X)$ of step functions is a subspace of the normed linear space of bounded functions $(L^\infty([a, b], X), \|\cdot\|_{L^\infty})$.

8.1.2 Multivariable, Banach-Valued Integration

Now we need to define the n -dimensional volume or *measure* of an n -interval. Note that we define the measure of an n -interval to be the same regardless of whether the boundary or part of the boundary of the interval is included or not.

Definition 8.1.6. For each $j \in \{1, \dots, n\}$ let $A_j \subset \mathbb{R}$ be an interval of the form $[a_j, b_j]$, (a_j, b_j) , $(a_j, b_j]$, or $[a_j, b_j)$. Let $R \subset \mathbb{R}^n$ be the Cartesian product $R = A_1 \times \dots \times A_n \subset \mathbb{R}^n$. Define the measure $\lambda(R)$ of the n -interval R to be the product of the lengths

$$\lambda(R) = \prod_{j=1}^n (b_j - a_j).$$

The definition of the integral of a step function is essentially identical to that of the one-dimensional case.

Definition 8.1.7. The integral of a step function $s = \sum_{I \in \mathcal{D}} \mathbf{x}_I \mathbb{1}_{R_I}$ is

$$\mathcal{J}(s) = \int_{[\mathbf{a}, \mathbf{b}]} s = \sum_{I \in \mathcal{D}} \mathbf{x}_I \lambda(R_I).$$

The proof of the next proposition is very similar to the single-variable case (Proposition 5.10.5).

Proposition 8.1.8. For any compact interval $[\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^n$, the integral operator $\mathcal{J} : S([\mathbf{a}, \mathbf{b}]; X) \rightarrow X$ is a bounded linear transformation with norm $\|\mathcal{J}\| = \lambda([\mathbf{b} - \mathbf{a}])$.

Proof. The proof is Exercise 8.2. $\square$

By the continuous linear extension theorem (Theorem 5.7.6), the integral extends from step functions to all functions in the closure (in the L^∞ -norm) of $S([\mathbf{a}, \mathbf{b}]; X)$. In particular, since the closure includes continuous functions $C([\mathbf{a}, \mathbf{b}]; X)$, this defines the integral for all continuous functions. The arguments in the proof of the single-variable Banach-valued integration theorem (Theorem 5.10.6) apply again in the multivariable case to give the following theorem.

Theorem 8.1.9 (Multivariable Banach-Valued Integral). Let $\overline{S([\mathbf{a}, \mathbf{b}]; X)}$ be the closure of $S([\mathbf{a}, \mathbf{b}]; X)$ in $L^\infty([\mathbf{a}, \mathbf{b}]; X)$. The linear transformation $\mathcal{J} : S([\mathbf{a}, \mathbf{b}]; X) \rightarrow X$ can be extended uniquely to a bounded linear transformation $\overline{\mathcal{J}} : \overline{S([\mathbf{a}, \mathbf{b}]; X)} \rightarrow X$, and

$$\|\overline{\mathcal{J}}\| = \lambda([\mathbf{b} - \mathbf{a}]).$$

Moreover, we have

$$C([\mathbf{a}, \mathbf{b}]; X) \subset \overline{S([\mathbf{a}, \mathbf{b}]; X)} \subset L^\infty([\mathbf{a}, \mathbf{b}]; X).$$

Proof. The proof is Exercise 8.3. $\square$

Definition 8.1.10. For any $[\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^n$ we denote the set $\overline{S([\mathbf{a}, \mathbf{b}]; X)}$ by $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$. The functions in $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$ are called regulated-integrable functions. For any $f \in \mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$, we call the linear transformation $\overline{\mathcal{J}}(f)$ in Theorem 8.1.9 the integral of f , and we usually denote it by

$$\int_{[\mathbf{a}, \mathbf{b}]} f = \overline{\mathcal{J}}(f).$$

Proposition 8.1.11. If $f, g \in \mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$, then the following hold:

- (i) $\|\int_{[\mathbf{a}, \mathbf{b}]} f\| \leq \lambda([\mathbf{a}, \mathbf{b}]) \sup_{\mathbf{t} \in [\mathbf{a}, \mathbf{b}]} \|f(\mathbf{t})\|$.
- (ii) For any sequence $(f_k)_{k=0}^\infty$ in $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$ that converges uniformly to f , we have

$$\lim_{k \rightarrow \infty} \int_{[\mathbf{a}, \mathbf{b}]} f_k = \int_{[\mathbf{a}, \mathbf{b}]} \lim_{k \rightarrow \infty} f_k = \int_{[\mathbf{a}, \mathbf{b}]} f.$$

(iii) Let $\|f\|$ denote the function $\mathbf{t} \mapsto \|f(\mathbf{t})\|$ from $[\mathbf{a}, \mathbf{b}]$ to $\mathbb{R}$. We have

$$\left\| \int_{[\mathbf{a}, \mathbf{b}]} f \right\| \leq \int_{[\mathbf{a}, \mathbf{b}]} \|f\|.$$

(iv) If $\|f(\mathbf{t})\| \leq \|g(\mathbf{t})\|$ for every $\mathbf{t} \in [\mathbf{a}, \mathbf{b}]$, then $\int_{[\mathbf{a}, \mathbf{b}]} \|f\| \leq \int_{[\mathbf{a}, \mathbf{b}]} \|g\|$.

Proof. The proof of (i) and (ii) are Exercise 8.5.

Item (iii) holds for step functions by the triangle inequality, and since every $f \in \mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$ is a uniform limit of step functions, we can use (ii) to conclude that (iii) holds for all $f \in \mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$.

Finally, if $h \in \mathcal{R}([\mathbf{a}, \mathbf{b}]; \mathbb{R})$ and if $h(\mathbf{t}) \geq 0$ for every $\mathbf{t} \in [\mathbf{a}, \mathbf{b}]$, there is a sequence of step functions $(s_k)_{k=0}^{\infty}$ that converges uniformly to h . Hence for any $\varepsilon > 0$ there is an $N > 0$ such that $\|s_k - h\|_{L^\infty} < \varepsilon/\lambda([\mathbf{a}, \mathbf{b}])$, whenever $k \geq N$. This implies that $s_k(\mathbf{t}) > -\varepsilon/\lambda([\mathbf{a}, \mathbf{b}])$ for every $\mathbf{t} \in [\mathbf{a}, \mathbf{b}]$ whenever $k \geq N$, and thus

$$\int_{[\mathbf{a}, \mathbf{b}]} s_k \geq -\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we have

$$\int_{[\mathbf{a}, \mathbf{b}]} h = \lim_{k \rightarrow \infty} \int_{[\mathbf{a}, \mathbf{b}]} s_k \geq 0.$$

Letting $h = \|g\| - \|f\|$ gives (iv). $\square$

Remark 8.1.12. As in the single-variable case, one can easily check that the Riemann construction of the integral defines a bounded linear transformation on $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$ that agrees with our definition on step functions, and hence by the uniqueness part of the continuous linear extension theorem must agree with our construction on all of $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$. The Riemann construction does work for a slightly larger space of functions than $\mathcal{R}([\mathbf{a}, \mathbf{b}]; X)$; however, we need the integral to be defined on yet more functions, because in applications we must often move limits past integrals, but many limits of Riemann-integrable functions are not Riemann integrable. This is discussed in more depth in Section 8.2.

8.1.3 Integration over subsets of $[\mathbf{a}, \mathbf{b}]$

We have defined integration over intervals $[\mathbf{a}, \mathbf{b}]$, but we would like to define it over a more general set $E \subset [\mathbf{a}, \mathbf{b}]$. To do this, we *extend by zero* using the indicator function (see Definition 8.1.4) to make a function (denoted $f\mathbb{1}_E$) on all of $[\mathbf{a}, \mathbf{b}]$.

Definition 8.1.13. For any function $f : E \rightarrow X$, the extension of f by zero is the function

$$f\mathbb{1}_E(z) = \begin{cases} f(z), & z \in E, \\ 0, & z \notin E. \end{cases}$$

The obvious way to try to define $\int_E f$ for $f : E \rightarrow X$ is to extend f by zero and then integrate over the whole interval to get $\int_{[a,b]} \mathbb{1}_E f$. This definition of integration allows us to define integrals of many functions over many different sets, but there are some important problems. The most immediate problem is the fact that there are many sets for which the indicator function itself is not in $\mathcal{R}([a, b]; \mathbb{R})$.

Unexample 8.1.14. Even the set consisting of a single point $p \in [a, b] \subset \mathbb{R}$ has an indicator function $\mathbb{1}_p$ that is not in $\mathcal{R}([a, b]; \mathbb{R})$. To see this, note first that step functions on any interval $[a, b] \subset \mathbb{R}$ are all right continuous, meaning that

$$\lim_{t \rightarrow t_0^+} s(t) = s(t_0)$$

for every $t_0 \in [a, b]$. Moreover, by Exercise 8.4 the uniform limit of a sequence of right-continuous functions is right continuous, and thus every function in $\mathcal{R}([a, b]; \mathbb{R})$ is right continuous. But $\mathbb{1}_p$ is not right continuous, so $\mathbb{1}_p \notin \mathcal{R}([a, b]; \mathbb{R})$.

There are many important sets E and functions f for which we would expect to be able to define the integral $\int_E f$, but for which the regulated integral (and also the more traditional Riemann construction) does not work. In the next section we discuss this problem and its solution in more depth.

8.2 Overview of Daniell–Lebesgue Integration

8.2.1 The Problem

The main problem with the integral, as we have defined it, is that we have not defined it for enough functions. The space of functions that we know how to integrate so far is the space $\mathcal{R}([a, b]; X)$, which includes continuous functions, but does not include any function that is not right continuous (see Unexample 8.1.14).

By definition, $\mathcal{R}([a, b]; X)$ is closed in the L^∞ -norm, and it is a subspace of $L^\infty([a, b]; X)$, which is complete by Theorem 5.7.5, so $\mathcal{R}([a, b]; X)$ is also complete. This means that any Cauchy sequence $(f_n)_{n=0}^\infty$ in $\mathcal{R}([a, b]; X)$ converges uniformly to some function $f \in \mathcal{R}([a, b]; X)$, and since integration is continuous with respect to this norm, we have

$$\lim_{n \rightarrow \infty} \int_{[a,b]} f_n = \int_{[a,b]} \lim_{n \rightarrow \infty} f_n = \int_{[a,b]} f.$$

This is an extremely useful property, but uniform convergence is a very strong property—too strong for many of the applications we are interested in.

We need similar results for sequences and limits that may not converge uniformly, but rather only converge in the L^1 -norm.

Definition 8.2.1. For any $f \in \mathcal{R}([a, b]; X)$, define $\|f\|_{L^1}$ to be

$$\|f\|_{L^1} = \int_{[a,b]} \|f\|.$$

We call $\|\cdot\|_{L^1}$ the L^1 -norm.

Proposition 8.2.2. *The function $\|\cdot\|_{L^1}$ is a norm on $\mathcal{R}([a, b]; X)$.*

Proof. The proof is Exercise 8.6. $\square$

Nota Bene 8.2.3. Although we previously defined the integral for Banach-valued functions, and although most of the results of this chapter hold for general Banach-valued functions, for the rest of the chapter we restrict ourselves to $\mathbb{R}$ -valued functions (that is, to the case of $X = \mathbb{R}$) just to keep things simple.

Note that the case of real-valued functions can easily be used to describe the case where f takes values in any finite-dimensional Banach space. For a complex-valued function $f = u + iv$, simply define the integral to be the sum of two real-valued integrals

$$\int_{[a,b]} f = \int_{[a,b]} u + i \int_{[a,b]} v,$$

and for $X = \mathbb{F}^n$ write $f = (f_1, \dots, f_n)$ and then define

$$\int_{[a,b]} f = \left(\int_{[a,b]} f_1, \dots, \int_{[a,b]} f_n \right).$$

Integration should be a continuous linear transformation on some space of functions V that is complete in the L^1 -norm. So, if a sequence $(g_n)_{n=0}^\infty$ in V is Cauchy in the L^1 -norm, then we need $\lim_{n \rightarrow \infty} g_n$ to exist in V , and the limit of integrals $\lim_{n \rightarrow \infty} \int_{[a,b]} g_n$ ought to be equal to $\int_{[a,b]} \lim_{n \rightarrow \infty} g_n$. But, unfortunately, $\mathcal{R}([a, b]; \mathbb{R})$ is not complete in the L^1 -norm, and neither is the space of Riemann-integrable functions. The solution is to construct a larger space of functions and extend the integral to that larger space.

In this section we briefly sketch the following main ideas:

- (i) How to construct a vector space $L^1([a, b]; \mathbb{R})$ containing both $S([a, b]; \mathbb{R})$ and $\mathcal{R}([a, b]; \mathbb{R})$ as subspaces.
- (ii) How to define integration and the L^1 -norm on $L^1([a, b]; \mathbb{R})$ in such a way that
 - (a) the new definition of integration agrees with the existing definition of integration for $\mathcal{R}([a, b]; \mathbb{R})$;
 - (b) the new definition of integration is a bounded (continuous) linear transformation with respect to the L^1 -norm;
 - (c) the space $L^1([a, b]; \mathbb{R})$ is complete with respect to the L^1 -norm.

The full proofs of these constructions and their properties are given in Chapter 9.

Nota Bene 8.2.4. If $X = \mathbb{R}$, then the integral, as we have defined it here, is often called the *Daniell integral*. We will usually call it the *Lebesgue integral*, however, because it is equivalent to the Lebesgue integral, and that name is more familiar to most mathematicians.

It is also common to use the phrase *Lebesgue integration* when talking about integration of functions in $L^1([a, b]; \mathbb{R})$, but it is important to note that there is really just one “integration” going on here. In particular, $\mathcal{R}([a, b]; \mathbb{R}) \subset L^1([a, b]; \mathbb{R})$, and Lebesgue integration restricted to $\mathcal{R}([a, b]; \mathbb{R})$ is just what you have always called integration since your first introduction to calculus. That is, Lebesgue integration is essentially just a way to extend regular old Riemann integration to a much larger collection of functions.

8.2.2 Sketch of Daniell–Lebesgue Integration

The basic strategy for constructing the space $L^1([a, b]; \mathbb{R})$ is first to show that *every* normed linear space S can be embedded uniquely into a complete normed linear space $\widehat{S}$ (called the *completion* of S) in such a way that S is dense in $\widehat{S}$, and then to show that when $S = \mathcal{R}([a, b]; \mathbb{R})$, the elements of $L^1([a, b]; \mathbb{R}) = \widehat{S}$ correspond to well-defined functions. The details of this are given in Section 9.1, but we give an overview of the main ideas here.

Given a normed linear space $(S, \|\cdot\|)$, we construct $\widehat{S}$ in two steps (for details see Section 9.1). First, let

$$S' = \{(s_n)_{n=0}^\infty \mid (s_n)_{n=0}^\infty \text{ is Cauchy in } S\}$$

be the set of Cauchy sequences in S . The set S' is a vector space with vector addition and scalar multiplication defined as

$$\alpha(s_n)_{n=0}^\infty + \beta(t_n)_{n=0}^\infty = (\alpha s_n + \beta t_n)_{n=0}^\infty.$$

Define $\|\cdot\|_{S'}$ on S' by

$$\|(s_n)_{n=0}^\infty\|_{S'} = \lim_{n \rightarrow \infty} \|s_n\|.$$

Unfortunately, $\|\cdot\|_{S'}$ is not a norm because there are nonzero elements $(s_n)_{n=0}^\infty$ of S' that have $\|(s_n)_{n=0}^\infty\|_{S'} = 0$. But the set

$$K = \{(s_n)_{n=0}^\infty \in S' : \|(s_n)_{n=0}^\infty\|_{S'} = 0\}$$

is a vector subspace of S' . It is not hard to prove that $\|\cdot\|_{S'}$ defines a norm $\|\cdot\|_{\widehat{S}}$ on the quotient space

$$\widehat{S} = S'/K.$$

The elements of $\widehat{S}$ are equivalence classes of L^1 -Cauchy sequences, where any two sequences $(s_n)_{n=0}^\infty$ and $(t_n)_{n=0}^\infty$ are equivalent if $\|s_n - t_n\| \rightarrow 0$. We prove in Section 9.1 that $(\widehat{S}, \|\cdot\|_{\widehat{S}})$ is complete and that S can be mapped injectively into $\widehat{S}$ by sending the element $s \in S$ to the constant sequence $(s_n = s \text{ for all } n \in \mathbb{N})$, and the two norms on S clearly agree (that is, $\|s\|_S = \|s\|_{\widehat{S}}$). Moreover, the subspace S is dense in $\widehat{S}$.

We define the space $L^1([a, b]; \mathbb{R})$ to be the completion of $S([a, b]; \mathbb{R})$ in the L^1 -norm; that is,

$$L^1([a, b]; \mathbb{R}) = \widehat{S}([a, b]; \mathbb{R}).$$

This is guaranteed to be complete, but we have two wrinkles to iron out. The first wrinkle arises from the way the completion is defined. We want $L^1([a, b]; \mathbb{R}) = \widehat{S}([a, b]; \mathbb{R})$ to consist of functions, not equivalence classes of sequences of functions. So, for each element of $L^1([a, b]; \mathbb{R})$, that is, for each equivalence class of L^1 -Cauchy sequences, we must give a well-defined function. To do this, we take the pointwise limit of a Cauchy sequence in the equivalence class. We prove in Section 9.3 that for each equivalence class there is at least one sequence that converges pointwise to a function. So we have an associated function arising from each element of $L^1([a, b]; \mathbb{R})$.

But the second wrinkle is that two different sequences in the same equivalence class can converge pointwise to different functions (see Unexample 8.2.5). Thus, unfortunately, there is not a *well-defined* function for each element of $L^1([a, b]; \mathbb{R})$. One instance of this is given in Unexample 8.2.5.

Unexample 8.2.5. Let $s_n \in S([0, 1]; \mathbb{R})$ be the characteristic function of the box $[0, \frac{1}{2^n})$, and let $(t_n)_{n=0}^\infty$ be the zero sequence ($t_n = 0$ for all $n \in \mathbb{N}$). The sequence $(s_n)_{n=0}^\infty$ converges pointwise to the characteristic function of the singleton set $\{0\}$, but the sequence of L^1 -norms $\|s_n\|_{L^1}$ converges to 0. Integrating gives $\|s_n - t_n\|_{L^1} \rightarrow 0$, and so these two Cauchy sequences are in the same equivalence class in $L^1([0, 1]; \mathbb{R})$, but their pointwise limits are not the same.

In a strict sense, there is no way to *solve* this problem—there is always more than one function that can be associated to a given element of $L^1([a, b]; \mathbb{R})$. But the two functions f and g associated to the same element of $L^1([a, b]; \mathbb{R})$ differ only on an infinitesimal set—a set of *measure zero*. We give a careful definition of measure zero in the next section.

We iron out the second wrinkle by treating two functions as being “the same” if they differ only on a set of measure zero; that is, we define an equivalence relation on the set of all functions by saying two functions are equivalent if they differ on a set of measure zero. In other words, f and g are equivalent if $f - g$ is supported on a set of measure zero. In this case we say f and g are *equal almost everywhere*, and we write $f = g$ a.e. This allows us to associate a unique equivalence class of functions to each element of $L^1([a, b]; \mathbb{R})$.

The upshot of all this is that we have constructed a vector space $L^1([a, b]; \mathbb{R})$ of functions (or rather a vector space of equivalence classes of functions) and a norm $\|\cdot\|_{L^1}$ on $L^1([a, b]; \mathbb{R})$ such that $L^1([a, b]; \mathbb{R})$ is complete with respect to this norm. Moreover, the set $S([a, b]; \mathbb{R})$ of step functions is a dense subspace of $L^1([a, b]; \mathbb{R})$, and our same old definition of integration of step functions is still a bounded linear transformation on $S([a, b]; \mathbb{R})$ with respect to the L^1 -norm. So the continuous linear extension theorem guarantees there is a unique way to extend integration to give

a bounded linear operator on all of $L^1([a, b]; \mathbb{R})$. This is the *Daniell* or *Lebesgue* integral.

Definition 8.2.6. We say that a function $f: [a, b] \rightarrow \mathbb{R}$ is integrable on $[a, b]$ if $f \in L^1([a, b]; \mathbb{R})$, that is, if f is almost everywhere equal to the pointwise limit of an L^1 -Cauchy sequence of step functions on $[a, b]$.

Nota Bene 8.2.7. Beware that although it is equivalent to the more traditional definition of $L^1([a, b]; \mathbb{R})$, our definition of $L^1([a, b]; \mathbb{R})$ is very different from the definition that you would see in a standard course on integration. In most treatments of integration, $L^1([a, b]; \mathbb{R})$ is defined to be the set of (equivalence classes a.e. of) measurable functions on $[a, b]$ for which $\|f\|_{L^1}$ is finite. But for us, $L^1([a, b]; \mathbb{R})$ is the set of (equivalence classes a.e. of) functions that are almost everywhere equal to the pointwise limit of an L^1 -Cauchy sequence of step functions on $[a, b]$.

Finally, the following proposition and its corollaries show that every sequence that is Cauchy with respect to the L^∞ -norm is also Cauchy with respect to the L^1 -norm; so $\mathcal{R}([a, b]; \mathbb{R})$ is a subspace of $L^1([a, b]; \mathbb{R})$, and the new definition of integration, when restricted to $\mathcal{R}([a, b]; \mathbb{R})$, agrees with our earlier definition of integration.

Proposition 8.2.8. For any $f \in \mathcal{R}([a, b]; \mathbb{R})$ we have

$$\|f\|_{L^1} \leq \lambda([a, b])\|f\|_{L^\infty}. \quad (8.1)$$

Proof. The proof is Exercise 8.7. $\square$

Corollary 8.2.9. If $(f_n)_{n=0}^\infty$ is a sequence in $\mathcal{R}([a, b]; \mathbb{R})$ converging uniformly to f , then $(f_n)_{n=0}^\infty$ also converges to f in the L^1 -norm.

Proof. The proof is Exercise 8.8 $\square$

Corollary 8.2.10. For any normed linear space X , if $T: \mathcal{R}([a, b]; \mathbb{R}) \rightarrow X$ (or $T: S([a, b]; \mathbb{R}) \rightarrow X$) is a bounded linear transformation with respect to the L^1 -norm, then it is also a bounded linear transformation with respect to the L^∞ -norm.

Proof. We have

$$\|T\|_{L^\infty, X} = \sup_f \frac{\|T(f)\|_X}{\|f\|_{L^\infty}} \leq \sup_f \frac{\lambda([a, b])\|T(f)\|_X}{\|f\|_{L^1}} \leq \lambda([a, b])\|T\|_{L^1, X},$$

where the suprema are both taken over all $f \in \mathcal{R}([a, b]; \mathbb{R})$. Thus, if $\|T\|_{L^1, X} < \infty$, then we also have $\|T\|_{L^\infty, X} < \infty$. $\square$

Corollary 8.2.11. *Let $\mathcal{J} : \mathcal{R}([a, b]; \mathbb{R}) \rightarrow \mathbb{R}$ be a linear transformation that, when restricted to step functions $S([a, b]; \mathbb{R})$, agrees with the integral $\mathcal{I} : S([a, b]; \mathbb{R}) \rightarrow \mathbb{R}$. If $\mathcal{J}$ is a bounded transformation with respect to the L^1 -norm, then $\mathcal{J}$ must be equal to the integral $\mathcal{I} : \mathcal{R}([a, b]; \mathbb{R}) \rightarrow \mathbb{R}$ defined in Theorem 8.1.9.*

Proof. The proof is Exercise 8.10. $\square$

8.3 Measure Zero and Measurability

In practical terms, the main new idea in our new theory of integration is the idea of *sets of measure zero*. It should not be surprising that some sets are so small as to have no volume. In Definition 8.1.6, for example, we defined an open interval $(a, b) = (a_1, b_1) \times \cdots \times (a_n, b_n)$ to have the same measure as a closed interval $[a, b]$. This would suggest that the measure of $[a, b] \setminus (a, b)$ should be 0.

It turns out that extending the definition of measure from intervals to more general sets is a subtle thing, so we do not treat it here (for a careful treatment of measure theory, see Volume 3), but defining a set of measure zero is not nearly so hard. In this section we define sets of measure zero and also *measurable sets*, which are, essentially, the sets where it makes sense to talk about integration.

8.3.1 Sets of Measure Zero

To motivate the idea of sets of measure zero, consider two basic properties that one might expect from the definition of a *measure* λ (or volume) of a set, assuming λ is defined for all the sets involved. The first property we expect is that the measure of a set should be at least as big as the measure of any subset:

$$\text{If } B \subset A, \text{ then } \lambda(B) \leq \lambda(A).$$

This suggests that for any set A of measure zero, any subset $B \subset A$ should also have measure zero, and if $(C_k)_{k=0}^{\infty}$ is a sequence of sets whose measure goes to zero ($\lambda(C_k) \rightarrow 0$ as $k \rightarrow \infty$), then $\bigcap_{k \in \mathbb{N}} C_k$ should have measure zero.

The second property we expect is that the measure of a union of sets should be no bigger than the sum of the measures of the individual pieces:

$$\lambda(A \cup B) \leq \lambda(A) + \lambda(B).$$

This suggests that any finite or countable union of sets of measure zero should also have measure zero.

Unfortunately, we only know how to define the measure of intervals so far—not more general sets—but these two properties suggest how to define sets of measure zero using only intervals.

Definition 8.3.1. *A set $A \subset \mathbb{R}^n$ has measure zero if for any $\varepsilon > 0$ there is a countable collection of n -intervals $(I_k)_{k=0}^{\infty}$, such that $A \subset \bigcup_{k=0}^{\infty} I_k$ and $\sum_{k=0}^{\infty} \lambda(I_k) < \varepsilon$.*

Proposition 8.3.2. *The following hold:*

- (i) *Any subset of a set of measure zero has measure zero.*

- (ii) A single point in $\mathbb{R}^n$ has measure zero.
- (iii) A countable union of sets of measure zero has measure zero.

Proof. Item (i) follows immediately from the definition. The proof of (ii) is Exercise 8.11.

For (iii) assume that $(C_k)_{k \in \mathbb{N}}$ is a countable collection of sets of measure zero. Assume that $\varepsilon > 0$ is given. For each $k \in \mathbb{N}$ there exists a collection $(I_{j,k})_{j \in \mathbb{N}}$ of intervals covering C_k such that $\sum_{j \in \mathbb{N}} \lambda(I_{j,k}) < \varepsilon/2^{k+1}$. Allowing k also to vary, the collection $(I_{j,k})_{j \in \mathbb{N}, k \in \mathbb{N}}$ of all these intervals is a countable union of countable sets, and hence is countable. Moreover, we have $\bigcup_{k \in \mathbb{N}} C_k \subset \bigcup_{j,k \in \mathbb{N}} I_{j,k}$ and

$$\sum_{k \in \mathbb{N}} \sum_{j \in \mathbb{N}} \lambda(I_{j,k}) < \sum_{k \in \mathbb{N}} \varepsilon/2^{k+1} = \varepsilon. \quad \square$$

Example 8.3.3. The *Cantor ternary set* is constructed by starting with the closed interval $C_0 = [0, 1] \subset \mathbb{R}$ and removing the (open) middle third to get $C_1 = [0, 1/3] \cup [2/3, 1]$. Repeating the process, removing the middle third of each of the preceding collection of intervals, gives $C_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$. Continuing this process gives a sequence of sets $(C_k)_{k=0}^\infty$, such that each C_k consists of 2^k closed intervals of total length $(2/3)^k$.

Define the Cantor ternary set to be the intersection $C_\infty = \bigcap_{k \in \mathbb{N}} C_k$. Since $[0, 1]$ is compact and each C_k is closed, C_∞ is nonempty (see Theorem 5.5.11). One can actually show that this set is uncountable (see, for example, [Cha95, Chap. 1, Sect. 4.4]).

To see that C_∞ has measure zero, note that each C_k contains C_∞ , and C_k is a finite union of closed intervals of total length $(2/3)^k$. Since $(2/3)^k$ can be made arbitrarily small by choosing large enough values of k , the set C_∞ satisfies the conditions of Definition 8.3.1.

Definition 8.3.4. We say that functions f and g are equal almost everywhere and write $f = g$ a.e. if the set $\{t \mid f(t) \neq g(t)\}$ has measure zero in $\mathbb{R}^n$.

Example 8.3.5. Integration gives the same result for any two functions that are equal almost everywhere, so when we need to integrate a function that is messy or otherwise difficult work with, but all the “bad” parts of the function are supported on a set of measure zero, we can replace it with a function that is equal to the first function almost everywhere, but is (hopefully) easier to integrate.

For example, the *Dirichlet function* defined on $\mathbb{R}$ by

$$f(t) = \begin{cases} 1 & \text{if } t \text{ is rational,} \\ 0 & \text{if } t \text{ is irrational} \end{cases}$$

is equal to zero almost everywhere, because the set where $f(t) \neq 0$ is countable, and hence has measure zero. Since $f = 0$ a.e., for any interval $[a, b] \subset \mathbb{R}$ we have $\int_a^b f = \int_a^b 0 = 0$.

Proposition 8.3.6. *The relation $=$ a.e. defines an equivalence relation on the set of all functions from $[a, b]$ to $\mathbb{R}$.*

Proof. The proof is Exercise 8.12. $\square$

Definition 8.3.7. *We say that a sequence of functions $(f_k)_{k=0}^\infty$ converges almost everywhere if the set $\{t \mid (f_k(t))_{k=0}^\infty \text{ does not converge}\}$ has measure zero. If for almost all t the sequence $(f_k(t))_{k=0}^\infty$ converges to $f(t)$, we write $f_k \rightarrow f$ a.e.*

Note that *convergence almost everywhere* is about pointwise convergence. It does not depend on a choice of norm (for example, the L^∞ - or L^1 -norms) on the space of functions.

Example 8.3.8. Consider the sequence of functions given by

$$f_n = n \mathbb{1}_{[0, \frac{1}{n}]}$$

The sequence $(f_n(0))_{n=0}^\infty$ does not converge, but for all $x \neq 0$ the sequence $(f_n(x))_{n=0}^\infty$ converges to zero. Hence, $f_n(x) \rightarrow 0$ a.e.

8.3.2 Measurability

We often wish to integrate functions over a set that is not a compact interval. Unfortunately, we cannot do this consistently for all sets and all functions. The functions for which we can even think of defining a sensible integral are called *measurable functions*, and the sets where we can sensibly talk about the possibility of integrating functions are called *measurable sets*.

Definition 8.3.9. *A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is measurable if there is a sequence of step functions $(s_k)_{k=0}^\infty$ (not necessarily L^1 -Cauchy) such that $s_k \rightarrow f$ a.e. We say that a set $A \subset \mathbb{R}^n$ is measurable if its indicator function $\mathbb{1}_A$ (see Definition 8.1.4) is measurable.*

We now define integration on measurable sets that are contained in some compact interval, that is, on bounded measurable sets.

Definition 8.3.10. *If $A \subset [a, b]$ is measurable and $f \mathbb{1}_A \in L^1([a, b]; \mathbb{R})$, then we define*

$$\int_A f = \int_{[a, b]} f \mathbb{1}_A.$$

Define $L^1(A; \mathbb{R})$ to be the set of functions $f: A \rightarrow \mathbb{R}$ such that $f\mathbb{1}_A \in L^1([a, b]; \mathbb{R})$.

The next proposition and its corollary show that the integral over A is well defined; that is, it is independent of the choice of interval $[a, b]$.

Proposition 8.3.11. *Let $A \subset [a, b] \subset [c, d]$ be measurable. We have $f\mathbb{1}_A \in L^1([a, b]; \mathbb{R})$ if and only if $f\mathbb{1}_A \in L^1([c, d]; \mathbb{R})$. Moreover,*

$$\int_{[a, b]} f\mathbb{1}_A = \int_{[c, d]} f\mathbb{1}_A. \quad (8.2)$$

Proof. ($\implies$) If $f\mathbb{1}_A \in L^1([a, b]; \mathbb{R})$, then there is a sequence $(s_n)_{n=0}^\infty$ of step functions on $[a, b]$ such that $(s_n)_{n=0}^\infty$ is L^1 -Cauchy on $[a, b]$ and $s_n \rightarrow f\mathbb{1}_A$ a.e. Extending each of these step functions by zero outside of $[a, b]$ gives step functions t_m on $[c, d]$, and $t_m \rightarrow f\mathbb{1}_A$ a.e.

From the definition of the integral of a step function, we have

$$\int_{[c, d]} |t_n - t_m| = \int_{[a, b]} |s_n - s_m| \quad \text{and} \quad \int_{[c, d]} t_n = \int_{[a, b]} s_n$$

for all $n, m \in \mathbb{N}$. Hence, $(t_n)_{n=0}^\infty$ is L^1 -Cauchy on $[c, d]$, the function $f\mathbb{1}_A \in L^1([c, d]; \mathbb{R})$, and (8.2) holds.

($\impliedby$) If $f\mathbb{1}_A \in L^1([c, d]; \mathbb{R})$, then there is a sequence $(t_n)_{n=0}^\infty$ of step functions on $[c, d]$ such that $(t_n)_{n=0}^\infty$ is L^1 -Cauchy on $[c, d]$ and $t_n \rightarrow f\mathbb{1}_A$ a.e. Multiplying each of these by $\mathbb{1}_{[a, b]}$ gives step functions $s_n = t_n\mathbb{1}_{[a, b]}$.

Now we show that $(s_n)_{n=0}^\infty$ is L^1 -Cauchy on $[a, b]$. Given $\varepsilon > 0$, choose $N > 0$ such that $\|t_n - t_m\|_{L^1} < \varepsilon$ (on $[c, d]$) whenever $n, m > N$. The L^1 -norm of $s_n - s_m$ on $[a, b]$ is

$$\int_{[a, b]} |s_n - s_m| = \int_{[a, b]} |t_n - t_m|\mathbb{1}_{[a, b]} \leq \int_{[c, d]} |t_n - t_m| < \varepsilon.$$

Thus, $(s_m)_{m=0}^\infty$ is L^1 -Cauchy on $[a, b]$, and hence $f\mathbb{1}_A \in L^1([a, b]; \mathbb{R})$.

Finally, $(s_n)_{n=0}^\infty$ also defines an L^1 -Cauchy sequence on $[c, d]$ converging to $f\mathbb{1}_A$ a.e. with $\int_{[a, b]} s_n = \int_{[c, d]} s_n$ for every n . Therefore $(s_n)_{n=0}^\infty$ and $(t_n)_{n=0}^\infty$ define the same element of $L^1([c, d]; \mathbb{R})$, and they have the same integral. Hence, (8.2) holds. $\square$

Since any nonempty intersection of two compact intervals is again a compact interval, we get the following corollary.

Corollary 8.3.12. *If $A \subset [a, b] \cap [a', b']$ is measurable, then $f\mathbb{1}_A \in L^1([a, b]; \mathbb{R})$ if and only if $f\mathbb{1}_A \in L^1([a', b']; \mathbb{R})$. Moreover,*

$$\int_{[a, b]} f\mathbb{1}_A = \int_{[a', b']} f\mathbb{1}_A. \quad (8.3)$$

Nota Bene 8.3.13. Most sets that you are likely to encounter in applied mathematics are measurable, including all open and closed sets and any countable unions and intersections of open or closed sets. But not every subset of $\mathbb{R}^n$ is measurable. We do not provide an example here, but you can find examples in [Van08, RF10].

8.4 Monotone Convergence and Integration on Unbounded Domains

The main result of this section is the *monotone convergence theorem*, which is the first of three very important convergence theorems. Since $L^1([a, b]; \mathbb{R})$ is complete (by definition), any L^1 -Cauchy sequence converges in the L^1 -norm to a function in $L^1([a, b]; \mathbb{R})$. For such a sequence $(f_k)_{k=0}^\infty$ the limit commutes with the integral

$$\lim_{k \rightarrow \infty} \int_{[a, b]} f_k = \int_{[a, b]} \lim_{k \rightarrow \infty} f_k.$$

But a sequence of functions that converges pointwise does not necessarily converge in the L^1 -norm (see Unexample 8.4.3), and its pointwise limit is not always integrable. The three convergence theorems give us conditions for identifying when a pointwise-convergent sequence is actually L^1 -Cauchy.

After discussing some basic integral properties, we state and prove the monotone convergence theorem. We conclude the section with an important consequence of the monotone convergence theorem, namely, integration on unbounded domains.

8.4.1 Some Basic Integral Properties

Definition 8.4.1. For any set A and any function $f: A \rightarrow \mathbb{R}$, define

$$f^+(a) = \begin{cases} f(a) & \text{if } f(a) \geq 0, \\ 0 & \text{if } f(a) \leq 0 \end{cases} \quad \text{and} \quad f^-(a) = \begin{cases} -f(a) & \text{if } f(a) \leq 0, \\ 0 & \text{if } f(a) \geq 0. \end{cases}$$

Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

The integral operator on $L^1([a, b]; \mathbb{R})$ is linear and continuous because it is the unique continuous linear extension of the integral operator on step functions, and continuity implies that it commutes with limits in the L^1 -norm. In Section 9.3 we also show that it satisfies the following basic properties.

Proposition 8.4.2. For any $f, g \in L^1([a, b]; \mathbb{R})$ we have the following:

- (i) If $f \leq g$ a.e., then $\int_{[a, b]} f \leq \int_{[a, b]} g$.
- (ii) $|\int_{[a, b]} f| \leq \int_{[a, b]} |f| = \|f\|_{L^1}$.
- (iii) The functions $\max(f, g)$, $\min(f, g)$, f^+ , f^- , and $|f|$ are all integrable.

- (iv) If $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a measurable function (see Definition 8.3.9), and if $|h| \in L^1([a, b]; \mathbb{R})$, then $h \in L^1([a, b]; \mathbb{R})$.
- (v) If $\|g\|_{L^\infty} \leq M < \infty$, then $fg \in L^1([a, b]; \mathbb{R})$ and $\|fg\|_{L^1} \leq M\|f\|_{L^1}$.

8.4.2 Monotone Convergence

As mentioned above, not every pointwise-convergent sequence of integrable functions converges in the L^1 -norm. This means that we cannot move limits past integral signs for these sequences, and in fact we cannot even expect that the limiting function is integrable.

Unexample 8.4.3. In general we cannot expect to be able to interchange the limit and the integral. Consider the sequence $(f_k)_{k=0}^\infty$ in $L^1([0, 1]; \mathbb{R})$ given by

$$f_k(x) = \begin{cases} 2^k, & x \in (0, 2^{-k}], \\ 0 & \text{otherwise.} \end{cases}$$

This sequence converges pointwise to the zero function, but $\int_{[0,1]} f_k = 1$ for all $k \in \mathbb{N}$, so

$$\lim_{k \rightarrow \infty} \int_{[0,1]} f_k = 1 \neq 0 = \int_{[0,1]} \lim_{k \rightarrow \infty} f_k.$$

In the rest of this section we discuss the *monotone convergence theorem*, which guarantees that if the integrals of a monotone sequence are bounded, then the sequence must be L^1 -Cauchy.

Definition 8.4.4. We say that a sequence of functions $(f_k)_{k=0}^\infty$ from $[a, b]$ to $\mathbb{R}$ is monotone increasing if for every $\mathbf{x} \in [a, b]$ we have $f_k(\mathbf{x}) \leq f_{k+1}(\mathbf{x})$ for every $k \in \mathbb{N}$. This is denoted $f_k \leq f_{k+1}$. We say that the sequence is almost everywhere monotone increasing if for every $k \in \mathbb{N}$ the set $\{\mathbf{x} \in [a, b] \mid f_k(\mathbf{x}) > f_{k+1}(\mathbf{x})\}$ has measure zero. This is denoted $f_k \leq f_{k+1}$ a.e. Monotone decreasing and almost everywhere monotone decreasing are defined analogously.

Theorem 8.4.5 (Monotone Convergence Theorem). Let $(f_k)_{k=0}^\infty \subset L^1([a, b]; \mathbb{R})$ be almost everywhere monotone increasing. If there exists $M \in \mathbb{R}$ such that

$$\int_{[a,b]} f_k \leq M \tag{8.4}$$

for all $k \in \mathbb{N}$, then $(f_k)_{k=0}^\infty$ is L^1 -Cauchy, and hence there exists a function $f \in L^1([a, b]; \mathbb{R})$ such that

- (i) $f = \lim_{k \rightarrow \infty} f_k$ a.e. and
- (ii) $\int_{[a,b]} f = \int_{[a,b]} \lim_{k \rightarrow \infty} f_k = \lim_{k \rightarrow \infty} \int_{[a,b]} f_k$.

The same conclusion holds if $(f_k)_{k=0}^\infty \subset L^1([a, b]; \mathbb{R})$ is almost everywhere monotone decreasing on $\mathbb{R}^n$ and there exists $M \in \mathbb{R}$ such that

$$\int_{[a, b]} f_k \geq M. \quad (8.5)$$

This theorem is sometimes called the *Beppo Levi theorem*.

Proof. Monotonicity and Proposition 8.4.2(i) guarantee that for every $k \in \mathbb{N}$ we have $\int_{[a, b]} f_k \leq \int_{[a, b]} f_{k+1} \leq M$, so the sequence of real numbers $(\int_{[a, b]} f_k)_{k=0}^\infty$ is monotone increasing in $\mathbb{R}$ and bounded, and hence has a limit L .

For any $\varepsilon > 0$ there exists an $N > 0$ such that

$$0 < \left(L - \int_{[a, b]} f_k \right) < \varepsilon$$

whenever $k \geq N$. Choosing $\ell > m \geq N$, we have

$$\begin{aligned} \|f_\ell - f_m\|_{L^1} &= \int_{[a, b]} |f_\ell - f_m| = \int_{[a, b]} (f_\ell - f_m) \\ &= \left(L - \int_{[a, b]} f_m \right) - \left(L - \int_{[a, b]} f_\ell \right) \\ &< \varepsilon. \end{aligned}$$

Thus $(f_k)_{k=0}^\infty$ is L^1 -Cauchy, as required. The monotone decreasing case follows immediately from the previous result by replacing each f_k by $-f_k$ and replacing M by $-M$. $\square$

Remark 8.4.6. Notice that the sequence in Unexample 8.4.3 is not monotone increasing, so the theorem does not apply to that sequence.

8.4.3 Integration on Unbounded Domains

So far we have only defined integration over measurable subsets of bounded intervals. But here we use the monotone convergence theorem to extend the definition of integration (and integrability) to unbounded intervals.

Definition 8.4.7. Given an unbounded, measurable set $A \subset \mathbb{R}^n$ and a measurable, nonnegative function $f : A \rightarrow \mathbb{R}$, we say that f is integrable on A if there exists an increasing sequence $(E_k)_{k=0}^\infty$ (by increasing we mean each $E_k \subset E_{k+1}$) of bounded measurable subsets of A with $A = \bigcup_{k=0}^\infty E_k$, such that there exists $M \in \mathbb{R}$ with

$$\int_{E_k} f \leq M$$

for all k . Since the sequence of real numbers $\int_{E_k} f$ is nondecreasing and bounded, it must have a finite limit. Define the integral of f on A to be the limit

$$\int_A f = \lim_{k \rightarrow \infty} \int_{E_k} f.$$

We say that an arbitrary measurable function g is integrable on A if g^+ and g^- are both integrable on A , and we define

$$\int_A g = \int_A g^+ - \int_A g^-.$$

We write $L^1(A; \mathbb{R})$ to denote the set of equivalence classes of integrable functions on A (modulo equality almost everywhere).

Nota Bene 8.4.8. Exercise 8.18 shows that $L^1(A; \mathbb{R})$ is a normed linear space with the L^1 -norm and the very important property that a function g is integrable on A if and only if $|g|$ is integrable on A .

Example 8.4.9. If f is a nonnegative continuous function on $\mathbb{R}$, then it is integrable on every compact interval of the form $[0, n]$. If $\lim_{n \rightarrow \infty} \int_0^n f(x) dx = L$ is finite, then every $\int_0^n f(x) dx$ is bounded above by L and $\int_0^\infty f(x) dx = L$.

For example,

$$\int_0^\infty e^{-x} dx = \lim_{n \rightarrow \infty} \int_0^n e^{-x} dx = \lim_{n \rightarrow \infty} (1 - e^{-n}) = 1.$$

Unexample 8.4.10. The continuous function $f(x) = (1+x)/(1+x^2)$ is not integrable on all of $\mathbb{R}$ because

$$f^+(x) = \begin{cases} f(x), & x \geq -1, \\ 0, & x \leq -1, \end{cases}$$

has an integral that is unbounded on the intervals $[-n, n]$; that is,

$$\begin{aligned} \int_{[-n, n]} f^+ &= \int_{-1}^n (1+x)/(1+x^2) dx \\ &= \left[\operatorname{Arctan}(x) + \frac{1}{2} \log(1+x^2) \right] \Big|_{-1}^n \\ &= \operatorname{Arctan}(n) + \frac{1}{2} \log(1+n^2) + \pi/4 - \frac{1}{2} \log(2) \rightarrow \infty \quad \text{as } n \rightarrow \infty. \end{aligned}$$

This shows f^+ is not integrable on $\mathbb{R}$, and hence f is not integrable on $\mathbb{R}$.

If f^+ and f^- are not integrated separately, they may partially cancel each other. This is a problem because the cancellation may be different, depending on the increasing sequence of bounded subsets we use in the integration. For example, using intervals of the form $[-n, n]$ gives

$$\int_{[-n,n]} f = \int_{-n}^n (1+x)/(1+x^2) dx = 2 \operatorname{Arctan}(n) \rightarrow \pi \quad \text{as } n \rightarrow \infty.$$

But using the intervals of the form $[-n, n^2]$ gives

$$\int_{[-n,n^2]} f dx = \operatorname{Arctan}(n^2) + \operatorname{Arctan}(n) + \frac{1}{2} \log \left(\frac{1+n^4}{1+n^2} \right) \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

This shows that nonnegativity of the integrand is important to have consistent results.

Below we show that the definition of integrable and the value of the integral on unbounded domains do not depend on the choice of the sequence $(E_k)_{k=0}^\infty$ of measurable subsets, that is, they are well defined.

Theorem 8.4.11. *Let $A \subset \mathbb{R}^n$ be a measurable set, and let f be a nonnegative measurable function on A . Let $(E_k)_{k=0}^\infty$ and $(E'_k)_{k=0}^\infty$ be two increasing sequences of bounded measurable subsets such that*

$$A = \bigcup_{k=0}^{\infty} E_k = \bigcup_{k=0}^{\infty} E'_k.$$

If there exists an $M \in \mathbb{R}$ such that for all k we have $\int_{E_k} f \leq M$, then we also have $\int_{E'_k} f \leq M$ for all k , and

$$\lim_{k \rightarrow \infty} \int_{E_k} f = \lim_{k \rightarrow \infty} \int_{E'_k} f.$$

Proof. Since each E_k is bounded and measurable, there exists a compact interval $[a, b]$ with $E_k \subset [a, b]$, and such that $\mathbb{1}_{E_k} \in L^1([a, b]; \mathbb{R})$. Similarly, for every m we may choose $[a', b']$ containing E'_m and such that $\mathbb{1}_{E'_m} \in L^1([a', b']; \mathbb{R})$. Let $[c, d]$ be a compact interval containing both $[a, b]$ and $[a', b']$. By Proposition 8.3.11 we have that $\mathbb{1}_{E_k}$ and $\mathbb{1}_{E'_m}$ are in $L^1([c, d]; \mathbb{R})$, and by Proposition 8.4.2(v) the products $\mathbb{1}_{E_k} \mathbb{1}_{E'_m}$ and $\mathbb{1}_{E_k} \mathbb{1}_{E'_m} f$ are in $L^1([c, d]; \mathbb{R})$. Therefore, the restriction of $\mathbb{1}_{E_k}$ and $\mathbb{1}_{E_k} f$ to E'_m both lie in $L^1(E'_m; \mathbb{R})$. Trading the roles of E'_m and E_k in the previous argument implies that $\mathbb{1}_{E'_k}$ and $\mathbb{1}_{E'_k} f$ are in $L^1(E_m; \mathbb{R})$.

If $\int_{E_k} f \leq M$ for all k , then

$$\int_{E'_k} f \mathbb{1}_{E_m} = \int_{E_m} f \mathbb{1}_{E'_k} \leq \int_{E_m} f \leq M$$

for all k and m . By the monotone convergence theorem, we have

$$\int_{E'_k} f = \int_{E'_k} \lim_{m \rightarrow \infty} f \mathbb{1}_{E_m} = \lim_{m \rightarrow \infty} \int_{E'_k} f \mathbb{1}_{E_m} \leq M. \quad (8.6)$$

Assume now that $L = \lim_{k \rightarrow \infty} \int_{E_k} f$ and $L' = \lim_{k \rightarrow \infty} \int_{E'_k} f$. Since the sequences $(\int_{E_k} f)_{k=0}^\infty$ and $(\int_{E'_k} f)_{k=0}^\infty$ are nondecreasing, they satisfy $\int_{E_k} f \leq L$ for all k and

$\int_{E'_k} f \leq L'$ for all k . Taking $M = L$ and taking the limit of (8.6) as $k \rightarrow \infty$ gives $L' \leq L$. Similarly, interchanging the roles of E_k and E'_k and setting $M = L'$ gives $L \leq L'$. Thus $L = L'$, as required. $\square$

8.5 Fatou's Lemma and the Dominated Convergence Theorem

The monotone convergence theorem is the key to proving two other important convergence theorems: *Fatou's lemma* and the *dominated convergence theorem*. Like the monotone convergence theorem, these two theorems give conditions that guarantee a pointwise-convergent sequence is actually L^1 -Cauchy.

Nota Bene 8.5.1. Being L^1 -Cauchy only guarantees convergence in the space $L^1([a, b]; \mathbb{R})$. In general an L^1 -Cauchy sequence of functions in $\mathcal{R}([a, b]; \mathbb{R})$ does not converge in the subspace $\mathcal{R}([a, b]; \mathbb{R})$ or in the space of Riemann integrable functions. That is, the monotone convergence theorem, Fatou's lemma, and the dominated convergence theorem are usually only useful if we work in $L^1([a, b]; \mathbb{R})$.

8.5.1 Fatou's Lemma

Recall that the *limit inferior* of a sequence $(x_k)_{k=0}^\infty$ of real numbers is defined as

$$\liminf_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} \left(\inf_{m \geq k} x_m \right),$$

and the *limit superior* is defined as

$$\limsup_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} \left(\sup_{m \geq k} x_m \right).$$

These always exist (although they may be infinite), even if the limit does not.

Fatou's lemma tells us when the $\liminf$ of a sequence (not necessarily convergent) of nonnegative integrable functions is integrable, and it tells us how the integral of the $\liminf$ is related to the $\liminf$ of the integrals. Fatou's lemma is also the key tool we use to prove the dominated convergence theorem.

Theorem 8.5.2 (Fatou's Lemma). *Let $(f_k)_{k=0}^\infty$ be a sequence of integrable functions on $[a, b]$ that are almost everywhere nonnegative, that is, for every $k \in \mathbb{N}$ we have $f_k(\mathbf{x}) \geq 0$ for for almost every $\mathbf{x}$. If*

$$\liminf_{k \rightarrow \infty} \int_{[a, b]} f_k < \infty,$$

then

- (i) $(\liminf_{k \rightarrow \infty} f_k) \in L^1([a, b]; \mathbb{R})$ and
- (ii) $\int_{[a, b]} \liminf_{k \rightarrow \infty} f_k \leq \liminf_{k \rightarrow \infty} \int_{[a, b]} f_k.$

Proof. First we show that the infimum of any sequence $(f_\ell)_{\ell=1}^\infty$ of almost-everywhere-nonnegative integrable functions must also be integrable. For each $k \in \mathbb{N}$ let g_k be the function defined by $g_k(t) = \min\{f_1(t), f_2(t), \dots, f_k(t)\}$. The sequence $(g_k)_{k=0}^\infty$ is a monotone decreasing sequence of almost-everywhere-nonnegative functions with $\lim_{k \rightarrow \infty} g_k = g = \inf_{m \in \mathbb{N}} f_m$. Since every g_k is almost everywhere nonnegative, we have $\int_{[a,b]} g_k \geq 0$. By the monotone convergence theorem (Theorem 8.4.5), we have $\inf_{m \in \mathbb{N}} \int_{[a,b]} f_m = \lim_{k \rightarrow \infty} \int_{[a,b]} g_k = \int_{[a,b]} g \in L^1([a,b]; \mathbb{R})$.

Now for each $k \in \mathbb{N}$ let $h_k = \inf_{\ell \geq k} f_\ell \in L^1([a,b]; \mathbb{R})$. Each h_k is almost everywhere nonnegative, and the sequence is monotone increasing, with $\lim_{k \rightarrow \infty} h_k = \liminf_{k \rightarrow \infty} f_k < \infty$. Moreover, for each $n \in \mathbb{N}$ we have $h_n \leq f_n$, so $\int_{[a,b]} h_n \leq \int_{[a,b]} f_n$ and, taking limits, we have

$$\int_{[a,b]} h_n \leq \lim_{k \rightarrow \infty} \int_{[a,b]} h_k = \liminf_{k \rightarrow \infty} \int_{[a,b]} h_k \leq \liminf_{k \rightarrow \infty} \int_{[a,b]} f_k < \infty.$$

Therefore the monotone convergence theorem guarantees $\liminf_{k \rightarrow \infty} h_k$ is integrable and $\lim_{k \rightarrow \infty} \int_{[a,b]} h_k = \int_{[a,b]} \liminf_{k \rightarrow \infty} h_k$. Combining the previous steps gives

$$\int_{[a,b]} \liminf_{k \rightarrow \infty} f_k = \int_{[a,b]} \lim_{k \rightarrow \infty} h_k = \lim_{k \rightarrow \infty} \int_{[a,b]} h_k \leq \liminf_{k \rightarrow \infty} \int_{[a,b]} f_k. \quad \square$$

Remark 8.5.3. The inequality in Fatou's lemma should not be surprising. For intuition about this, consider the situation where the sequence (f_k) consists of only two nonnegative functions f_0 and f_1 . This is depicted in Figure 8.4. In this case

$$\inf(f_0, f_1) = \min(f_0, f_1) \leq f_0 \quad \text{and} \quad \inf(f_0, f_1) = \min(f_0, f_1) \leq f_1,$$

so by Proposition 8.4.2(i), we have

$$\int_{[a,b]} \inf(f_0, f_1) \leq \int_{[a,b]} f_0 \quad \text{and} \quad \int_{[a,b]} \inf(f_0, f_1) \leq \int_{[a,b]} f_1.$$

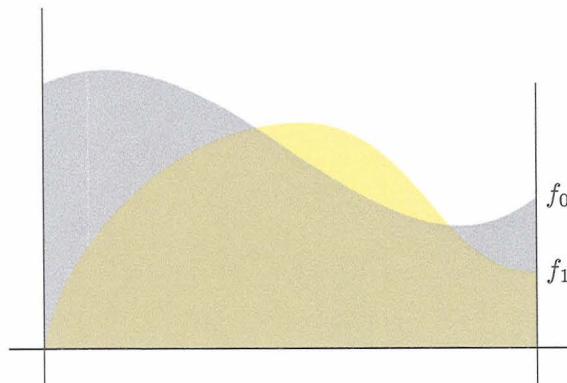


Figure 8.4. An elementary variant of the inequality in Fatou's lemma, as described in Remark 8.5.3. The area under $\min(f_0, f_1)$ (dark yellow intersection) is less than the area under f_0 (gray) and less than the area under f_1 (yellow). In symbols, this is $\int_a^b \min(f_0, f_1) \leq \min(\int_a^b f_0, \int_a^b f_1)$.

This implies

$$\int_{[a,b]} \inf(f_0, f_1) \leq \inf \left(\int_{[a,b]} f_0, \int_{[a,b]} f_1 \right).$$

It should be clear that the same is true for any finite collection of integrable functions. The inequality in Fatou's lemma is the analogous result for an infinite sequence of functions.

Example 8.5.4. The inequality in Fatou's lemma cannot be replaced by an equality. For each $k \in \mathbb{Z}^+$ let $f_k = k \mathbb{1}_{[0, 1/k]} \in L^1([0, 1]; \mathbb{R})$. We have

$$\int_0^1 \liminf_k (f_k) = \int_0^1 0 = 0 < 1 = \liminf_k 1 = \liminf_k \int_0^1 f_k$$

8.5.2 Dominated Convergence

The last of the three big convergence theorems is the *dominated convergence theorem*, which says that for any sequence of nonnegative functions that converges pointwise, if it is bounded above (dominated) by an integrable function almost everywhere, then it converges in the L^1 -norm.

Theorem 8.5.5 (Dominated Convergence Theorem). *Consider a sequence $(f_k)_{k=0}^\infty \subset L^1([a, b]; \mathbb{R})$ of integrable functions that converges pointwise almost everywhere to f . If there exists an integrable function $g \in L^1([a, b]; \mathbb{R})$ such that $|f_k| \leq g$ a.e. for every $k \in \mathbb{N}$, then $f \in L^1([a, b]; \mathbb{R})$ and*

$$\lim_{k \rightarrow \infty} \int_{[a,b]} f_k = \int_{[a,b]} \lim_{k \rightarrow \infty} f_k = \int_{[a,b]} f. \quad (8.7)$$

Proof. For each $k \in \mathbb{N}$ the function $h_k = g - f_k$ is nonnegative almost everywhere. Moreover,

$$\int_{[a,b]} h_k = \int_{[a,b]} g - f_k \leq \|g\|_{L^1} + \|f_k\|_{L^1} \leq 2\|g\|_{L^1},$$

so Fatou's lemma gives $g - \lim_{k \rightarrow \infty} f_k = \liminf_{k \rightarrow \infty} h_k \in L^1([a, b]; \mathbb{R})$, which implies that $f = \lim_{k \rightarrow \infty} f_k = g + \liminf_{k \rightarrow \infty} h_k \in L^1([a, b]; \mathbb{R})$. Moreover, we have

$$\begin{aligned} \int_{[a,b]} g - \int_{[a,b]} \lim_{k \rightarrow \infty} f_k &= \int_{[a,b]} \lim_{k \rightarrow \infty} h_k \\ &= \int_{[a,b]} \liminf_{k \rightarrow \infty} h_k \\ &\leq \limsup_{k \rightarrow \infty} \int_{[a,b]} h_k \\ &= \int_{[a,b]} g - \liminf_{k \rightarrow \infty} \int_{[a,b]} f_k, \end{aligned}$$

and thus we see

$$\int_{[a,b]} \lim_{k \rightarrow \infty} f_k \geq \liminf_{k \rightarrow \infty} \int_{[a,b]} f_k.$$

Repeating the previous argument with $\tilde{h}_k = g + f_k$ gives the other direction, which gives (8.7). $\square$

Remark 8.5.6. The dominated convergence theorem guarantees that we can interchange limits if we can find an integrable function that dominates the sequence almost everywhere. Note that in Unexample 8.4.3 no integrable function dominates all the terms f_k .

Example 8.5.7. For each $n \in \mathbb{N}$, let $f_n : \mathbb{R} \rightarrow \mathbb{R}$ be given by

$$f_n(x) = (1 - x/n)^n e^{x/3}.$$

We wish to evaluate

$$\lim_{n \rightarrow \infty} \int_0^T f_n dx = \lim_{n \rightarrow \infty} \int_0^T (1 - x/n)^n e^{x/3} dx$$

for arbitrary $T > 0$, but it is not immediately obvious how to integrate $\int_0^T f_n dx$.

Since $\lim_{n \rightarrow \infty} (1 - x/n)^n = e^{-x}$, the sequence $(f_n)_{n=0}^\infty$ converges pointwise to the function $f(x) = e^{-2x/3}$. It is easy to check that e^z is greater than or equal to $1+z$ for all $z \in \mathbb{R}$, and therefore $(1 - x/n) \leq e^{-x/n}$ for all x . This gives

$$f_n(x) = (1 - x/n)^n e^{x/3} \leq e^{-2x/3} = f(x),$$

and since f is continuous, it is integrable. Therefore the dominated convergence theorem applies and gives

$$\lim_{n \rightarrow \infty} \int_0^T f_n dx = \int_0^T \lim_{n \rightarrow \infty} f_n dx = \int_0^T f dx = \int_0^T e^{-2x/3} dx,$$

which is easy to integrate.

As a nice consequence of the dominated convergence theorem, we can now give a useful condition for deciding when an infinite sum of integrable functions converges and is integrable.

Proposition 8.5.8. If $(f_k)_{k=0}^\infty$ is any sequence of functions in $L^1([a, b]; \mathbb{R})$ such that

$$\sum_{k=0}^{\infty} \int_{[a,b]} |f_k| < \infty,$$

then $\sum_{k=0}^{\infty} f_k$ converges almost everywhere on $[a, b]$, and

$$\int_{[a,b]} \sum_{k=0}^{\infty} f_k = \sum_{k=0}^{\infty} \int_{[a,b]} f_k.$$

Proof. By Exercise 8.16 we have that for almost every $\mathbf{x} \in [a, b]$ the series $\sum_{k=0}^{\infty} |f_k(\mathbf{x})|$ converges, and the resulting function $\sum_{k=0}^{\infty} |f_k|$ is integrable. Therefore, for almost every $\mathbf{x} \in [a, b]$ the series $\sum_{k=0}^{\infty} f_k(\mathbf{x})$ converges absolutely, and hence it converges, by Proposition 5.6.13.

The partial sums of $\sum_{k=0}^{\infty} f_k$ are all dominated by the integrable function $\sum_{k=0}^{\infty} |f_k|$, so by the dominated convergence theorem the series $\sum_{k=0}^{\infty} f_k$ is integrable and

$$\int_{[a,b]} \sum_{k=0}^{\infty} f_k(x) = \sum_{k=0}^{\infty} \int_{[a,b]} f_k. \quad \square$$

8.6 Fubini's Theorem and Leibniz's Integral Rule

An important consequence of the monotone convergence theorem is *Fubini's theorem*, which allows us to convert multivariable integrals into repeated single-variable integrals, where we can use the fundamental theorem of calculus and other standard tools of single-variable integration. We note that Fubini's theorem is easy to prove for step functions, but integrating more general functions in $L^1([a, b]; \mathbb{R})$ involves taking limits of step functions, and iterated integrals involve taking two different limits. So it should not be too surprising that a convergence theorem is needed to control the way the two limits interact.

Finally, as an additional benefit, Fubini's theorem gives *Leibniz's integral rule*, which gives conditions for when we can differentiate under the integral sign. This turns out to be a very useful tool and can simplify many difficult problems.

8.6.1 Fubini's Theorem

Throughout this section we fix $X = [a, b] \subset \mathbb{R}^n$ and $Y = [c, d] \subset \mathbb{R}^m$. Since we integrate functions over each of these sets, in order to reduce confusion we write the symbols dx , dy , or $dx dy$ at the end of our integrals to help indicate where the integration is taking place, so we write $\int_X g(\mathbf{x}) dx$ to indicate the integral of $g \in L^1(X; \mathbb{R})$, and we write $\int_{X \times Y} f(\mathbf{x}, \mathbf{y}) dx dy$ to indicate the integral of $f \in L^1(X \times Y; \mathbb{R})$.

Theorem 8.6.1 (Fubini's Theorem). *Assume that $f: X \times Y \rightarrow \mathbb{R}$ is integrable on $X \times Y \subset \mathbb{R}^{m+n}$. For each $\mathbf{x} \in X$ consider the function $f_{\mathbf{x}}: Y \rightarrow \mathbb{R}$ given by $f_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$. We have the following:*

- (i) *For almost all $\mathbf{x} \in X$ the function $f_{\mathbf{x}}$ is integrable on Y .*
- (ii) *The function $F: X \rightarrow \mathbb{R}$ given by*

$$F(\mathbf{x}) = \begin{cases} \int_Y f_{\mathbf{x}}(\mathbf{y}) dy & \text{if } f_{\mathbf{x}} \text{ is integrable,} \\ 0 & \text{otherwise} \end{cases}$$

is integrable on X .

(iii) *The integral of f may be computed as*

$$\int_{X \times Y} f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}d\mathbf{y} = \int_X F(\mathbf{x}) \, d\mathbf{x}.$$

Remark 8.6.2. We often write $\int_X (\int_Y f_{\mathbf{x}}(\mathbf{y}) \, d\mathbf{y}) \, d\mathbf{x}$ instead of $\int_X F(\mathbf{x}) \, d\mathbf{x}$. We call this an *iterated integral*.

Nota Bene 8.6.3. The reader should beware that the obvious converse of Fubini is not true. Integrability of $f_{\mathbf{x}}$ and of $F = \int_Y f_{\mathbf{x}}(\mathbf{y}) \, d\mathbf{y}$ is not sufficient to guarantee that f is integrable on $X \times Y$. However, with some additional conditions one can sometimes still deduce the integrability of f . For more details see [Cha95, Sect. IV.5.11].

We prove Fubini's theorem in Section 9.4. In the rest of this section we focus on its implications and how to use it.

Fubini's theorem allows us to reduce higher-dimensional integrals down to a repeated application of one-dimensional integrals, and these can often be computed by standard techniques, such as the fundamental theorem of calculus (Theorem 6.5.4).

Example 8.6.4. We can compute the integral of $f(x, y) = \cos(x)y^2$ over $R = (0, \pi/2) \times (0, 1) \subset \mathbb{R}^2$ using Fubini's theorem:

$$\begin{aligned} \int_R \cos(x)y^2 \, dxdy &= \int_0^{\pi/2} \left(\int_0^1 \cos(x)y^2 \, dy \right) dx \\ &= \int_0^{\pi/2} \frac{\cos(x)}{3} \, dx = \frac{1}{3}. \end{aligned}$$

8.6.2 Interchanging the Order of Integration

Proposition 8.6.5. *If $f: X \times Y \rightarrow \mathbb{R}$ is integrable, then the function $\tilde{f}: Y \times X \rightarrow \mathbb{R}$, given by $\tilde{f}(\mathbf{y}, \mathbf{x}) = f(\mathbf{x}, \mathbf{y})$, is integrable, and*

$$\int_{X \times Y} f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}d\mathbf{y} = \int_{Y \times X} \tilde{f}(\mathbf{y}, \mathbf{x}) \, dyd\mathbf{x}.$$

Proof. The proof is Exercise 9.20. $\square$

Corollary 8.6.6. *If $f: X \times Y \rightarrow \mathbb{R}$ is integrable, then*

$$\int_X \left(\int_Y f_{\mathbf{x}}(\mathbf{y}) \, d\mathbf{y} \right) d\mathbf{x} = \int_{X \times Y} f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}d\mathbf{y} = \int_Y \left(\int_X f_{\mathbf{y}}(\mathbf{x}) \, d\mathbf{x} \right) d\mathbf{y}.$$

Proof. This follows immediately from Fubini's theorem and the proposition. $\square$

Corollary 8.6.6 is useful because changing the order of integration can often simplify a problem substantially.

Example 8.6.7. It is difficult to compute the inner antiderivative of the iterated integral

$$\int_{-\pi}^{\pi} \left(\int_1^2 \tan(3x^2 - x + 2) \sin(y) dx \right) dy.$$

But changing the order of integration gives

$$\int_{-\pi}^{\pi} \left(\int_1^2 \tan(3x^2 - x + 2) \sin(y) dx \right) dy = \int_1^2 \tan(3x^2 - x + 2) \left(\int_{-\pi}^{\pi} \sin(y) dy \right) dx$$

and $\int_{-\pi}^{\pi} \sin(y) dy = 0$, so the entire double integral is zero.

Example 8.6.8. Computing the iterated integral $\int_0^1 \int_x^1 e^{y^2} dy dx$ is not so easy, but notice that it is equal to $\int_0^1 \int_0^1 e^{y^2} \mathbb{1}_S(x, y) dy dx$, where S is the upper triangle obtained by cutting the unit square $(0, 1) \times (0, 1)$ in half, diagonally. Interchanging the order of integration, using Corollary 8.6.6, gives

$$\int_0^1 \int_x^1 e^{y^2} dy dx = \int_0^1 \int_0^1 e^{y^2} \mathbb{1}_S(x, y) dy dx = \int_0^1 \int_0^1 e^{y^2} \mathbb{1}_S(x, y) dx dy.$$

Now for each value of y , we have $(x, y) \in S$ if and only if $0 \leq x \leq y$, so the integral becomes

$$\int_0^1 \int_0^y e^{y^2} dx dy,$$

which is easily evaluated to $\int_0^1 y e^{y^2} dy = \frac{1}{2}(e - 1)$.

8.6.3 Leibniz's Integral Rule

An important consequence of Fubini's theorem is Leibniz's integral rule, which gives conditions for when we can differentiate under the integral sign (that is, when derivatives commute with integrals). This is a very useful tool that greatly simplifies many difficult problems.

Theorem 8.6.9 (Leibniz's Integral Rule). *Let $X = (a, b) \subset \mathbb{R}$ be an open interval, let $Y = [c, d] \subset \mathbb{R}$ be a compact interval, and let $f: X \times Y \rightarrow \mathbb{R}$ be*

continuous. If f has continuous partial derivative $\frac{\partial f(x,y)}{\partial x}$ at each point $(x,y) \in X \times Y$, then the function

$$\psi(x) = \int_c^d f(x,y) dy$$

is differentiable at each point $x \in X$ and

$$\frac{d}{dx} \psi(x) = \int_c^d \frac{\partial f(x,y)}{\partial x} dy. \quad (8.8)$$

Proof. Fix some $x_0 \in X$. Using the second form of the fundamental theorem of calculus (6.16) and Fubini's theorem, we have

$$\begin{aligned} \psi(x) - \psi(x_0) &= \int_c^d (f(x,y) - f(x_0,y)) dy \\ &= \int_c^d \left(\int_{x_0}^x \frac{\partial f(z,y)}{\partial z} dz \right) dy \\ &= \int_{x_0}^x \left(\int_c^d \frac{\partial f(z,y)}{\partial z} dy \right) dz. \end{aligned}$$

Letting

$$g(z) = \int_c^d \frac{\partial f(z,y)}{\partial z} dy,$$

we have

$$\psi(x) - \psi(x_0) = \int_{x_0}^x g(z) dz. \quad (8.9)$$

Using the first form of the fundamental theorem of calculus (6.15) to differentiate (8.9) with respect to x gives

$$\frac{d}{dx} \psi(x) = g(x). \quad \square$$

Example 8.6.10. Consider the problem of finding the derivative of the map $F(x) = \int_0^1 (x^2 + t)^3 dt$. We can solve this without Leibniz's rule by first using standard antidifferentiation techniques:

$$\begin{aligned} F(x) &= \int_0^1 (x^2 + t)^3 dt \\ &= \left. \frac{(x^2 + t)^4}{4} \right|_0^1 \\ &= \frac{(x^2 + 1)^4}{4} - \frac{x^8}{4}. \end{aligned}$$

From this we find $F'(x) = 2x(x^2 + 1)^3 - 2x^7$.

Alternatively, Leibniz says we can differentiate under the integral sign:

$$\begin{aligned}
 F'(x) &= \int_0^1 \frac{\partial}{\partial x} (x^2 + t)^3 dt \\
 &= \int_0^1 3(x^2 + t)^2 (2x) dt \\
 &= (x^2 + t)^3 (2x) \Big|_0^1 \\
 &= 2x(x^2 + 1)^3 - 2x^7.
 \end{aligned}$$

Remark 8.6.11. In the previous example it is not hard to verify that the two answers agree, but in many situations it is much easier to differentiate under the integral sign than it is to integrate first and differentiate afterward.

We can use Theorem 8.6.9 and the chain rule to prove a more general result. This generalized Leibniz formula is important in the proof of Green's theorem (Theorem 10.5.15).

Corollary 8.6.12. *Let X and A be open intervals in $\mathbb{R}$, and let $f: X \times A \rightarrow \mathbb{R}$ be continuous with continuous partial derivative $\frac{\partial f}{\partial x}$ at each point of $X \times A$. If $a, b: X \rightarrow A$ are differentiable functions and $\psi(x) = \int_{a(x)}^{b(x)} f(x, t) dt$, then $\psi(x)$ is differentiable and*

$$\frac{d}{dx} \psi(x) = \int_{a(x)}^{b(x)} \frac{\partial f(x, t)}{\partial x} dt - a'(x)f(x, a(x)) + b'(x)f(x, b(x)). \quad (8.10)$$

Proof. The proof is Exercise 8.29. $\square$

Example 8.6.13. Let $F(x) = \int_{\sin(x)}^{\cos(x)} \text{Arctan}(x+t) dt$. To compute $F'(x)$ we may use Corollary 8.6.12:

$$\begin{aligned}
 F'(x) &= \frac{d}{dx} \int_{\sin(x)}^{\cos(x)} \text{Arctan}(x+t) dt \\
 &= \int_{\sin(x)}^{\cos(x)} \frac{1}{1+(x+t)^2} dt - \cos(x) \text{Arctan}(x+\sin(x)) \\
 &\quad - \sin(x) \text{Arctan}(x+\cos(x)) \\
 &= (1 - \sin(x)) \text{Arctan}(x+\cos(x)) - (1 + \cos(x)) \text{Arctan}(x+\sin(x)).
 \end{aligned}$$

Example 8.6.14. Generalizing the previous example, for any constants $r, s \in \mathbb{R}$ and any function of the form $G(x) = \int_{a(x)}^{b(x)} g(rx + st) dt$, we have

$$\begin{aligned}
 G'(x) &= \int_{a(x)}^{b(x)} rg'(rx + st) dt - a'(x)g(rx + sa(x)) + b'(x)g(rx + sb(x)) \\
 &= \frac{r}{s}g(rx + st) \Big|_{a(x)}^{b(x)} - a'(x)g(rx + sa(x)) + b'(x)g(rx + sb(x)) \\
 &= \frac{r}{s}g(rx + sb(x)) - \frac{r}{s}g(rx + sa(x)) - a'(x)g(rx + sa(x)) \\
 &\quad + b'(x)g(rx + sb(x)) \\
 &= \left(\frac{r}{s} + b'(x)\right)g(rx + sb(x)) - \left(\frac{r}{s} + a'(x)\right)g(rx + sa(x)).
 \end{aligned}$$

8.7 Change of Variables

In one dimension the chain rule gives the very useful change of variables formula (6.19):

$$\int_c^d f(g(s))g'(s) ds = \int_{g(c)}^{g(d)} f(u) du,$$

which is sometimes known as the *substitution formula* or *u-substitution*.

We now extend this to a change of variables formula for higher dimensions. Often an integral that is hard to compute in one coordinate system is much easier to compute in another coordinate system, and the change of variables formula is precisely what we need to move from one system to another.

8.7.1 Diffeomorphisms

The type of function we use for a change of variables is called a *diffeomorphism*.

Definition 8.7.1. Let U and V be open subsets of $\mathbb{R}^n$. We say that $\Psi : U \rightarrow V$ is a diffeomorphism if Ψ is a C^1 bijection such that Ψ^{-1} is also C^1 .

Remark 8.7.2. The composition of two diffeomorphisms is a diffeomorphism. If $\Psi : U \rightarrow V$ is a diffeomorphism, then the derivative $D\Psi$ must be invertible at every point $\mathbf{t} \in U$ because the chain rule gives

$$I(\mathbf{t}) = (DI)(\mathbf{t}) = D(\Psi^{-1} \circ \Psi)(\mathbf{t}) = D(\Psi^{-1})(\Psi(\mathbf{t}))D\Psi(\mathbf{t}),$$

where $I : U \rightarrow U$ is the identity. Therefore $D(\Psi^{-1})(\Psi(\mathbf{t})) = ((D\Psi)(\mathbf{t}))^{-1}$.

Example 8.7.3.

- (i) The map $f : (0, 1) \rightarrow (-1, 1)$ given by $f(x) = 2x - 1$ is a diffeomorphism because $f^{-1}(y) = (y + 1)/2$ is C^1 .
- (ii) The map $g : (-1, 1) \rightarrow \mathbb{R}$ given by $g(x) = x/(1 - x^2)$ has $Dg(x) = (1 + x^2)/(1 - x^2)^2$, which is strictly positive for all $x \in (-1, 1)$, so g is strictly increasing and, hence, is injective. Since g is continuous and since $\lim_{x \downarrow -1} g(x) = -\infty$ and $\lim_{x \uparrow 1} g(x) = \infty$, the map g is also surjective. Thus, g has an inverse, and the inverse function theorem (Theorem 7.4.8) guarantees that the inverse is C^1 . Therefore, g is a diffeomorphism.

Using the quadratic formula gives $g^{-1}(y) = (-1 + \sqrt{1 + y^2})/2y$, but for a more complicated function, the inverse cannot always be written in a simple closed form. Nevertheless, we can often show the inverse exists from general arguments, and the inverse function theorem can often guarantee that the inverse is C^1 , even if we don't know explicitly what the inverse function is.

- (iii) The map $f(x, y) = (\cos(x), \sin(y))$ from the open square $(0, \pi/2) \times (0, \pi/2)$ to the open unit square $(0, 1) \times (0, 1)$ is a diffeomorphism because it is injective ($\cos$ and $\sin$ are both injective on $(0, \pi/2)$) and

$$Df(x, y) = \begin{bmatrix} -\sin(x) & 0 \\ 0 & \cos(y) \end{bmatrix}$$

has a nonzero determinant for all $(x, y) \in (0, \pi/2) \times (0, \pi/2)$, so the inverse function theorem guarantees that f^{-1} is C^1 .

Unexample 8.7.4.

- (i) The map $g : (-1, 1) \mapsto (-1, 1)$ given by $g(x) = x^3$ is not a diffeomorphism. Although it is bijective with inverse $g^{-1}(y) = y^{1/3}$, the inverse function is not C^1 at 0.
- (ii) Let U be the punctured plane $U = \mathbb{R}^2 \setminus \{0\}$, and let $h : (0, \infty) \times \mathbb{R} \rightarrow U$ be given by $h(r, t) = (r \cos(t), r \sin(t))$. It is straightforward to check that h is surjective. The derivative is

$$Dh = \begin{bmatrix} \cos(t) & -r \sin(t) \\ \sin(t) & r \cos(t) \end{bmatrix},$$

which has determinant $r > 0$, so the derivative is always invertible. Therefore, the inverse function theorem guarantees that in a sufficiently small neighborhood of any point of U there is a C^1 inverse of h . But there is no single function h^{-1} defined on all of U because h is not injective. Thus, h is not a diffeomorphism.

8.7.2 The Change of Variables Formula

In the single-variable change of variables formula (6.19), if g is a diffeomorphism on an open set containing $[c, d]$, then the derivative g' is continuous and never zero, so it may not change sign on the interval $[c, d]$. If $g' < 0$ on (c, d) , then $g(d) < g(c)$, so

$$\begin{aligned}\int_{g([c, d])} f &= \int_{g(d)}^{g(c)} f = - \int_{g(c)}^{g(d)} f(\tau) d\tau \\ &= - \int_c^d f(g(s))g'(s) ds = \int_{[c, d]} f(g(s))|g'(s)| ds.\end{aligned}$$

If $g' > 0$ on (c, d) , then

$$\int_{g([c, d])} f = \int_{g(c)}^{g(d)} f(\tau) d\tau = \int_c^d f(g(s))g'(s) ds = \int_{[c, d]} f(g(s))|g'(s)| ds.$$

In either case, we may write this as

$$\int_{g([c, d])} f = \int_{[c, d]} (f \circ g)|g'|. \quad (8.11)$$

This is essentially the form of the change of variables formula in higher dimensions. The main theorem of this section is the following.

Theorem 8.7.5 (Change of Variables Theorem). *Let U and V be open subsets of $\mathbb{R}^n$, let $X \subset U$ be a measurable subset of $\mathbb{R}^n$, and let $\Psi : U \rightarrow V$ be a diffeomorphism. The set $Y = \Psi(X)$ is measurable, and if $f : Y \rightarrow \mathbb{R}$ is integrable, then $(f \circ \Psi)|\det(D\Psi)|$ is integrable on X and*

$$\int_Y f = \int_X (f \circ \Psi)|\det(D\Psi)|. \quad (8.12)$$

Remark 8.7.6. In the special case that $\Psi : U \rightarrow V$ is a linear transformation, then $D\Psi = \Psi$ and the change of variables formula says that $\int_Y f = |\det(\Psi)| \int_X (f \circ \Psi)$.

An especially important special case of this is when X is an interval $[a, b] \subset \mathbb{R}^n$ and $f = 1$. In this case (8.12) says that the volume of $Y = \Psi([a, b])$ is exactly the determinant of Ψ times the volume of $[a, b]$. This should not be surprising—the singular value decomposition says that Ψ can be written as $U\Sigma V^T$, where U and V are orthonormal. Orthonormal matrices are products of rigid rotations and reflections, so they should not change the volume at all, and Σ is diagonal, so it scales the i th standard basis vector by σ_i . This changes the volume by the product of these σ_i —that is, by the determinant of Σ , which is the absolute value of the determinant of Ψ .

We prove the change of variables theorem in Section 9.5. For the rest of this section we discuss some implications and examples.

Example 8.7.7. Usually either the geometry of the region or the structure of the integrand gives a hint about which diffeomorphism to use for change of variables. Consider the integral

$$\iint_R \sin\left(\frac{y-x}{y+x}\right),$$

where R is the trapezoid region with vertices $(1, 0)$, $(3, 0)$, $(0, 3)$, and $(0, 1)$. Without a change of variables, this is not so easy to compute. But the presence of the terms $x+y$ and $x-y$ in the integrand and the fact that two of the sides of the trapezoid are segments of the lines $x+y=1$ and $x+y=3$ suggest the change of variables $u = y+x$ and $v = y-x$. Writing x and y in terms of u and v gives $\Psi(u, v) = (\frac{1}{2}(u-v), \frac{1}{2}(u+v))$ with

$$D\Psi = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \quad \text{and} \quad |\det(D\Psi)| = \frac{1}{2}.$$

Applying the change of variables formula with Ψ yields

$$\frac{1}{2} \int_1^3 \int_{-u}^u \sin\left(\frac{v}{u}\right) dv du = 0.$$

Corollary 8.7.8. If $\Psi : U \rightarrow V$ is a diffeomorphism, then for any subset $E \subset U$ of measure zero, the image $\Psi(E)$ has measure zero.

Proof. If E has measure zero, then $\int_{\Psi(E)} 1 = \int_E |\det(D\Psi)| = \int_U \mathbb{1}_E |\det(D\Psi)| = 0$ because $\mathbb{1}_E = 0$ a.e. The corollary now follows from Exercise 8.20. $\square$

8.7.3 Polar Coordinates

One important application of Theorem 8.7.5 is integration in polar coordinates. Define $\Psi : (0, \infty) \times (0, 2\pi) \rightarrow \mathbb{R}^2$ by $\Psi(r, \theta) = (r \cos(\theta), r \sin(\theta))$. This is just the restriction of the map h of Unexample 8.7.4(ii) to the set $(0, \infty) \times (0, 2\pi)$.

The map Ψ is a diffeomorphism from $(0, 2\pi) \times (0, \infty)$ to $\mathbb{R}^2 \setminus \{(x, 0) \mid x \geq 0\}$ (the plane with the origin and positive x -axis removed). To see this, first verify that Ψ is bijective, so it has an inverse. Moreover, Ψ is C^1 , and since $\det(D\Psi) = r > 0$ is never zero, the inverse function theorem implies that the inverse must also be C^1 .

For any $A \subset (0, 2\pi) \times (0, \infty)$ if we let $B = \Psi(A) \subset \mathbb{R}^2$, then (8.12) gives

$$\iint_B f(x, y) dx dy = \int_B f = \int_A (f \circ \Psi) r = \iint_A f(r \cos(\theta), r \sin(\theta)) r dr d\theta. \quad (8.13)$$

Moreover, we can extend this relation to $[0, 2\pi] \times [0, \infty)$ because the rays defined by $\theta = 0$ and $\theta = 2\pi$ have measure zero and hence contribute nothing to the integral.

Example 8.7.9. Let A be the region $\{(r, \theta) \mid 0 \leq \theta \leq \pi/3, 0 \leq r \leq \sqrt{\sin(3\theta)}\}$, and let $B = \Psi(A)$ be the corresponding region in rectangular coordinate, as in Figure 8.5.

The area of B is

$$\begin{aligned} \int_B 1 &= \int_A r \, dr \, d\theta = \int_0^{\pi/3} \int_0^{\sqrt{\sin(3\theta)}} r \, dr \, d\theta \\ &= \int_0^{\pi/3} \frac{1}{2} \sin(3\theta) \, d\theta = \frac{1}{3}. \end{aligned}$$

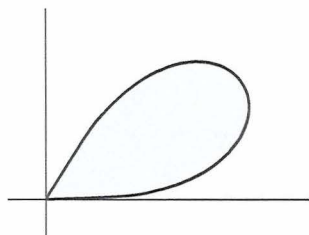


Figure 8.5. The petal-shaped region B of Example 8.7.9. Changing to polar coordinates simplifies the computation of the area of B .

Example 8.7.10. The integral $I = \int_{-\infty}^{\infty} e^{-x^2} \, dx$ plays an important role in probability and statistics, but it is not easy to integrate using traditional one-variable techniques. We can convert it to a two-dimensional integral that is easy to compute using polar coordinates, as follows:

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} \, dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} \, dy \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2-y^2} \, dx \, dy \\ &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)} \, dA = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r \, dr \, d\theta \\ &= \int_0^{2\pi} \frac{1}{2} \, d\theta = \pi. \end{aligned}$$

Thus $I = \sqrt{\pi}$.

Example 8.7.11. Sometimes more than one change of variable is needed. Consider the integral

$$\iint_E y^2 \, dA,$$

where E is the region bounded by the ellipse $16x^2 + 4y^2 = 64$. Since circles are generally easier to work with than ellipses, it is natural to make the substitution $\Psi_1(u, v) = (u, 2v)$, so that

$$D\Psi_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad |\det(D\Psi_1)| = 2.$$

This yields the integral

$$2 \iint_C 4v^2 dA,$$

where C is the region bounded by the circle $u^2 + v^2 = 4$. Switching to polar coordinates $\Psi_2(\theta, r) = (r \cos(\theta), r \sin(\theta))$ gives

$$8 \int_0^{2\pi} \int_0^2 r^3 \sin^2 \theta dr d\theta = 32\pi.$$

8.7.4 Spherical and Hyperspherical Coordinates

Spherical coordinates in $\mathbb{R}^3$ are similar to polar coordinates in $\mathbb{R}^2$. The three new coordinates are r, θ, ϕ , where θ is the angle in the xy -plane, ϕ is the angle from the z axis, and r is the radius, as depicted in Figure 8.6.

Definition 8.7.12. Let $U = (0, 2\pi) \times (0, \pi) \times (0, \infty)$, and define spherical coordinates $S : U \rightarrow \mathbb{R}^3$ by $S(\theta, \phi, r) = (r \sin(\phi) \cos(\theta), r \sin(\phi) \sin(\theta), r \cos(\phi))$.

We have

$$DS = \begin{bmatrix} -r \sin(\phi) \sin(\theta) & r \cos(\phi) \cos(\theta) & \sin(\phi) \cos(\theta) \\ r \sin(\phi) \cos(\theta) & r \cos(\phi) \sin(\theta) & \sin(\phi) \sin(\theta) \\ 0 & -r \sin(\phi) & \cos(\phi) \end{bmatrix}$$

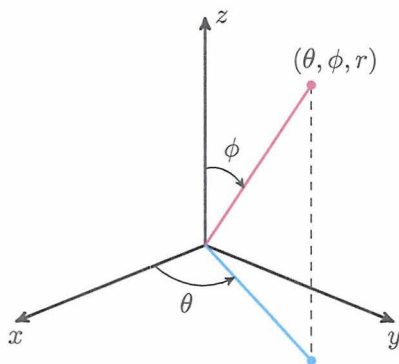


Figure 8.6. Representation of a point (red) in spherical coordinates (θ, ϕ, r) in 3-space. Observe that θ is the angle in the xy -plane to the projection (blue) of the point to that plane, whereas ϕ is the angle from the z -axis.

and

$$|\det(DS)| = r^2 \sin(\phi).$$

It is straightforward to check that S is C^1 and bijective onto $S(U) = V$, and hence has an inverse. Since $\det(DS)$ never vanishes, the inverse function theorem guarantees that the inverse is C^1 , and thus S is a diffeomorphism. The change of variables formula (8.12) gives

$$\begin{aligned} \int_{S(X)} f &= \iiint_{S(X)} f(x, y, z) dx dy dz = \int_X (f \circ S) r^2 \sin(\phi) \\ &= \int_X f(r \sin(\phi) \cos(\theta), r \sin(\phi) \sin(\theta), r \cos(\phi)) r^2 \sin(\phi) dr d\phi d\theta \end{aligned} \quad (8.14)$$

for any measurable $X \subset U$.

Example 8.7.13. Consider the region $D = [0, 2\pi] \times [0, \pi/6] \times [0, R]$, which when mapped by S gives an ice-cream-cone-shaped solid $C \subset \mathbb{R}^3$ as in Figure 8.7. As with polar coordinates, spherical coordinates are not bijective if we include the boundary, but the boundary has measure zero, so it contributes nothing to the integral.

Using (8.14) we see the volume of C is given by

$$\int_C 1 = \int_0^R \int_0^{\pi/6} \int_0^{2\pi} r^2 \sin(\phi) d\theta d\phi dr = \frac{(2 - \sqrt{3})\pi R^3}{3}.$$

The next definition generalizes polar and spherical coordinates to arbitrary dimensions as follows.

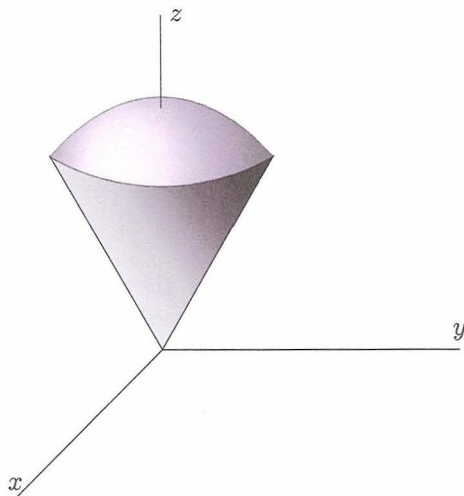


Figure 8.7. The ice-cream-cone-shaped region of Example 8.7.13.

Definition 8.7.14. Define hyperspherical coordinates on $\mathbb{R}^n$ by

$$\Psi : (0, \pi) \times \cdots \times (0, \pi) \times (0, 2\pi) \times (0, \infty) \rightarrow \mathbb{R}^n,$$

$$(\phi_1, \dots, \phi_{n-1}, r) \mapsto r \begin{bmatrix} \cos(\phi_1) \\ \sin(\phi_1) \cos(\phi_2) \\ \sin(\phi_1) \sin(\phi_2) \cos(\phi_3) \\ \vdots \\ \sin(\phi_1) \sin(\phi_2) \cdots \sin(\phi_{n-2}) \cos(\phi_{n-1}) \\ \sin(\phi_1) \sin(\phi_2) \cdots \sin(\phi_{n-2}) \sin(\phi_{n-1}) \end{bmatrix}.$$

A straightforward but tedious computation gives

$$\det(D\Psi) = r^{n-1} \sin^{n-2}(\phi_1) \sin^{n-3}(\phi_2) \cdots \sin(\phi_{n-2}),$$

so if we let $U = (0, \pi) \times \cdots \times (0, \pi) \times (0, 2\pi) \times (0, \infty)$ and $V = \Psi(U) \subset \mathbb{R}^n$, then

$$\int_V f = \int_U (f \circ \Psi) r^{n-1} \sin^{n-2}(\phi_1) \sin^{n-3}(\phi_2) \cdots \sin(\phi_{n-2}).$$

As with polar and spherical coordinates, one can check that Ψ is bijective to its image, and since $\det(D\Psi)$ does not vanish on its domain, the inverse function theorem guarantees that the inverse is C^1 , so Ψ is a diffeomorphism to its image. We use hyperspherical coordinates to find the volume of the unit ball in $\mathbb{R}^n$ (see Exercise 8.36).

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 8.1. Give an example of an interval $[a, b] \subset \mathbb{R}$ and a function $f: [a, b] \rightarrow \mathbb{R}$ that is not in $\mathcal{R}([a, b]; \mathbb{R})$, but whose absolute value $|f|$ is in $\mathcal{R}([a, b]; \mathbb{R})$.
- 8.2. Prove Proposition 8.1.8.

- 8.3. Prove Theorem 8.1.9.
 8.4. Prove the claim in Unexample 8.1.14 that the uniform limit of a sequence of right-continuous functions is right continuous.
 8.5. Prove (i) and (ii) of Proposition 8.1.11.

- 8.6. Prove Proposition 8.2.2, as follows:
 (i) First prove that $\|\cdot\|_{L^1}$ satisfies all the conditions of a norm (see Definition 3.5.1) except positivity.
 (ii) Prove that $\|f\|_{L^1} \geq 0$ for all $f \in \mathcal{R}([a, b]; X)$.
 (iii) If $f \in \mathcal{R}([a, b]; X)$ and if $f \neq 0$, then there exists some $\mathbf{p} \in [a, b]$ such that $\|f(\mathbf{p})\| > 0$. Prove that there exists an interval $R \subset [a, b]$ with $\lambda(R) > 0$ and with the property that for all $\mathbf{x} \in R$ we have $\|f(\mathbf{x})\| > \|f(\mathbf{p})\|/3$. Hint: By definition of $\mathcal{R}([a, b]; X)$ there exists a sequence $(s_n)_{n=0}^\infty$ of step functions that converges uniformly to f . Using the definition of uniform convergence, let $\varepsilon > 0$ be chosen so that $\varepsilon < \|f(\mathbf{p})\|/3$.
 (iv) Prove that $\int_{[a,b]} \|f\| \geq \int_R \|f\| \geq \lambda(R)\|f(\mathbf{p})\|/3 > 0$, and hence $\|f\|_{L^1} > 0$.
 8.7. Prove Proposition 8.2.8.
 8.8. Prove Corollary 8.2.9.
 8.9. Give an example of a sequence of step functions that is Cauchy in the L^1 -norm but is not Cauchy in the L^∞ -norm.
 8.10. Prove Corollary 8.2.11.

- 8.11. Prove that a single point in $\mathbb{R}^n$ has measure zero.
 8.12. Prove Proposition 8.3.6.
 8.13. Let $f_n(x) = n/(1 + n^2x^2)$. Show that $f_n \rightarrow 0$ a.e. as $n \rightarrow \infty$. What are all the values of x where $f_n(x) \not\rightarrow 0$?
 8.14. Prove that a bounded set $E \subset \mathbb{R}^n$ has measure zero if and only if $\mathbb{1}_E$ is integrable and $\int_E 1 = 0$.

- 8.15. For any $b > a > 0$, find (with proof) the value of $\lim_{n \rightarrow \infty} \int_a^b \log(x) e^{-nx} dx$.
 8.16.  Prove that if $(f_k)_{k=0}^\infty$ is any sequence of nonnegative functions in $L^1([a, b]; \mathbb{R})$ such that

$$\sum_{k=0}^{\infty} \int_{[a,b]} f_k < \infty,$$

then $\sum_{k=0}^{\infty} f_k$ converges almost everywhere on $[a, b]$, and

$$\int_{[a,b]} \sum_{k=0}^{\infty} f_k = \sum_{k=0}^{\infty} \int_{[a,b]} f_k.$$

Hint: Recall that $\sum_{k=0}^{\infty} f_k$ is $\lim_{N \rightarrow \infty} \sum_{k=0}^N f_k$.

- 8.17. Let $f \in L^1([a, b]; \mathbb{R})$ with $0 \leq f < 1$ a.e. Prove that

$$\lim_{k \rightarrow \infty} \int_{[a,b]} f^k = 0.$$

8.18. Let $A \subset \mathbb{R}^n$ be an unbounded measurable set. Prove that

- (i) a measurable function g is integrable on A if and only if $|g|$ is integrable on A ;
- (ii) the set $L^1(A; \mathbb{R})$ is a normed linear space with the L^1 -norm.

8.19. For each of the following functions, either prove that the function is integrable on $\mathbb{R}$ or prove that it is not integrable on $\mathbb{R}$. Hint: Some of the integrals cannot be evaluated easily, but you can still bound them using the basic properties of Proposition 8.4.2.

- (i) $f(x) = e^x \sin(e^x)$.
- (ii) $f(x) = e^x / (1 + e^{2x})$.
- (iii) $f(x) = \begin{cases} \sqrt{x}e^{-x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$
- (iv) $f(x) = \cos(x)/(1 + x^2)$.
- (v) $f(x) = \begin{cases} x^{-p}, & x \geq 1, \\ 0, & x < 1. \end{cases}$

(Decide and prove which values of p make f integrable and which do not.)

8.20. Prove the result of Exercise 8.14 in the case that E is not necessarily bounded.

8.21. Prove the *reverse Fatou lemma*: If $(f_k)_{k=0}^\infty$ is a sequence in $L^1([a, b]; \mathbb{R})$ and if there exists some nonnegative $g \in L^1([a, b]; \mathbb{R})$ with $f_k \leq g$ a.e. for all $k \in \mathbb{N}$, then

$$\limsup_{k \rightarrow \infty} \int_{[a, b]} f_k \leq \int_{[a, b]} \limsup_{k \rightarrow \infty} f_k.$$

8.22. Prove that in Fatou's lemma the condition that each f_k be nonnegative can be replaced with the condition that there exists some function $g \in L^1([a, b]; \mathbb{R})$ such that for all $k \in \mathbb{N}$ we have $f_k \geq g$ a.e.

8.23. Let

$$g(t) = \int_0^1 \frac{e^{-t^2(1+y^2)}}{1+y^2} dy.$$

Prove that $\lim_{t \rightarrow \infty} g(t) = 0$.

8.24. Let

$$f_n(x) = n/(1 + n^2 x^2).$$

Recall from Exercise 8.13 that $f_n \rightarrow 0$ a.e.

- (i) What does Fatou's lemma tell us about $\liminf_{n \rightarrow \infty} \int_a^b f_n dx$ in this case?
- (ii) Use one of the convergence theorems to prove that $\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = 0$ for all $b > a > 0$.
- (iii) Antidifferentiate explicitly to show that if $0 < b$, then $\lim_{n \rightarrow \infty} \int_0^b f_n(x) dx = \pi/2$, and if $a < 0 < b$, then $\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \pi$. Therefore,

$$\int_a^b \lim_{n \rightarrow \infty} f_n(x) dx = 0 < \pi/2 \leq \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx$$

when $a \leq 0$.

- (iv) Graph $f_n(x)$ in an open neighborhood of $x = 0$ for several values of $n \geq 1$.
- (v) Explain why both the monotone and dominated convergence theorems fail to apply for $a \leq 0 < b$.
- 8.25. Let $(f_k)_{k=0}^\infty$ be a sequence in $L^1([a, b]; \mathbb{R})$ with $f_k \rightarrow f$ a.e. Prove that if $|f| \leq g$ a.e. for some $g \in L^1([a, b]; \mathbb{R})$, then $f \in L^1([a, b]; \mathbb{R})$.

- 8.26. Let I be the integral $\int_C (2xy + 1) dx dy$, where C is the region bounded by $y = x^2$ and $y = 2x$. Write I as an iterated integral in two different ways. Compute the value of I .
- 8.27. Determine the volume of the region in $\mathbb{R}^n$ defined by the inequality

$$|x_1| + |x_2| + \cdots + |x_{n-1}| + |x_n| \leq r.$$

Justify (prove) your result. Hint: Consider the cases $n = 1$, $n = 2$, and $n = 3$. Once you find the volume in these three cases, you should be able to determine the pattern and use induction to prove it.

- 8.28. For every nonnegative integer n , let $\overline{B}(0, R) \subset \mathbb{R}^n$ be the closed n -ball of radius R . Prove that $\text{vol}(\overline{B}(0, R)) = R^n \text{vol}(\overline{B}(0, 1))$. Hint: Induct on n . For the induction step, slice the ball perpendicular to one axis to get an $(n - 1)$ -ball, and then integrate over all the slices.
- 8.29. Use Leibniz's integral rule (Theorem 8.6.9) and the chain rule (Theorem 6.4.7) to prove Corollary 8.6.12. Hint: Consider the function $\Phi(x, \alpha, \beta) = \int_\alpha^\beta f(x, t) dt$.
- 8.30. Show that $\int_0^\infty t^n e^{-t} dt = n!$ as follows:
- (i) For any $N > 0$ show that

$$\int_0^N e^{-tx} dt = \frac{1 - e^{-Nx}}{x}. \quad (8.15)$$

- (ii) Differentiate (8.15) repeatedly to show $\int_0^N (-1)^n t^n e^{-tx} dt = \frac{d^n}{dx^n} \frac{1 - e^{-Nx}}{x}$.
- (iii) Use the results of Exercise 6.26 to show for any $x > 0$ and for any $n \in \mathbb{N}$ that $\int_0^\infty t^n e^{-tx} dt = \frac{n!}{x^{n+1}}$.
- (iv) Evaluate at $x = 1$ to conclude that $\int_0^\infty t^n e^{-t} dt = n!$.
- 8.31. Show that $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$ as follows. Let $f(t) = (\int_0^t e^{-x^2} dx)^2$ and let $g(t) = \int_0^1 \frac{e^{-t^2(1+y^2)}}{1+y^2} dy$.
- (i) Show that $f'(t) + g'(t) = 0$ for all $t > 0$. Hint: After using Leibniz, consider the substitution $u = ty$.
- (ii) Show that $f(t) + g(t) = \pi/4$ for all $t > 0$.
- (iii) Use the result of Exercise 8.23 to show that $\lim_{t \rightarrow \infty} f(t) = \pi/4$ and hence that $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$.

- 8.32. Let $D \subset \mathbb{R}^2$ be the square with vertices $(2, 2)$, $(3, 3)$, $(2, 4)$, $(1, 3)$. Compute the integral

$$\int_D \ln(y^2 - x^2) dx dy.$$

8.33. Derive a formula for the volume of the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1.$$

More generally, if $A \in M_3(\mathbb{R})$ is positive definite, derive a formula for the volume of the ellipsoid $\mathbf{x}^\top A \mathbf{x} \leq 1$.

8.34. Given $A \in M_n(\mathbb{R})$ with $A > 0$, generalize Example 8.7.10 by showing that

$$\int_{\mathbb{R}^n} e^{-\mathbf{x}^\top A \mathbf{x}} d\mathbf{x} = \frac{\pi^{n/2}}{\sqrt{\det(A)}}.$$

8.35. Let Q be the open first quadrant $Q = \{(x, y) \mid x > 0, y > 0\}$, and let H be the upper half plane $H = \{(s, t) \mid t > 0\}$. Define $\Phi : Q \rightarrow H$ by $\Phi(x, y) = (x^2 - y^2, xy)$ for (x, y) in Q . For a point (x, y) in Q , the pair of numbers $(x^2 - y^2, xy) = \Phi(x, y)$ are called *hyperbolic coordinates* for (x, y) .

- (i) Show that $\Phi : Q \rightarrow \mathbb{R}^2$ is a diffeomorphism.
- (ii) Now define $D = \{(x, y) \mid x > 0, y > 0, 1 \leq x^2 - y^2 \leq 9, 2 \leq xy \leq 4\}$. Use hyperbolic coordinates to show that

$$\int_D (x^2 + y^2) dx dy = 8.$$

8.36. Let R denote the box $[0, \pi] \times \cdots \times [0, \pi] \times [0, 2\pi]$.

- (i) Use hyperspherical coordinates to express the volume of the unit n -ball in terms of the integral $I = \int_R \sin(\phi_1) \sin^2(\phi_2) \cdots \sin^{n-2}(\phi_{n-2})$.
- (ii) Use hyperspherical coordinates to express the integral $\int_{\mathbb{R}^n} e^{-\|\mathbf{x}\|^2} d\mathbf{x}$ in terms of I and the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.
- (iii) Use these results, combined with Example 8.7.10, to give a formula for the volume of the unit n -ball.
- (iv) Combine this with Exercise 8.28 to give a formula for the volume of any n -ball of radius r . (Alternatively, you could slightly generalize your computation in (i) to achieve the same result.)

Notes

Much of our treatment of integration in this chapter and the next is inspired by Soo Bong Chae's beautiful book [Cha95], which develops Lebesgue integration using an approach due to Riesz. Another source for the Riesz approach is [Soh14]. The Riesz approach has some significant similarities to our approach, but at heart they are still very different ways of looking at integration. Sources on the Daniell integral include [BM14], [AB66], and [Roy63] (the first edition of [RF10]). The Bochner integral is described in [Mik14]. Sources for a more standard approach to Lebesgue integration and measure theory include [Bre08, Jon93, RF10, Rud87].

Exercise 8.30 comes from Keith Conrad's "blurb" on differentiation under the integral sign [Con16]. Exercise 8.31 comes from Luc Rey-Bellet [Rey06]. Exercise 8.35 comes from Fitzpatrick [Fit06].

9

* Integration II

Nature laughs at the difficulties of integration.
—Pierre-Simon Laplace

In this chapter we give the details and remaining proofs of the development of the Daniell–Lebesgue integral, as outlined in the previous chapter.

9.1 Every Normed Space Has a Unique Completion

In this section we prove that every normed linear space can be embedded (uniquely) as a dense subspace of some Banach space. This is very handy in many settings, but our immediate interest is to use it to complete the space of step functions with respect to the L^1 -norm.

9.1.1 Seminorms and Norms

Before we begin our discussion of completion, we need one simple fact about seminorms. Recall from Definition 3.5.1 that seminorms are functions that are almost norms, but where $\|f\| = 0$ need not imply that $f = 0$.

Proposition 9.1.1. *If V is a vector space and $\|\cdot\|$ is a seminorm on V , then the set $K = \{\mathbf{v} \in V \mid \|\mathbf{v}\| = 0\}$ forms a vector subspace of V . Moreover, $\|\cdot\| : V/K \rightarrow \mathbb{R}$, defined by the rule $\|\mathbf{v} + K\| = \|\mathbf{v}\|$, forms a norm on the quotient space V/K .*

Proof. The proof is Exercise 9.1 $\square$

9.1.2 Every Normed Linear Space Has a Completion

The main theorem of this section is the following.

Theorem 9.1.2. *For any normed linear space $(X, \|\cdot\|)$, there exists a Banach space $(\widehat{X}, \|\cdot\|_{\widehat{X}})$ and an injective linear map $\phi : X \rightarrow \widehat{X}$ such that for every $\mathbf{x} \in X$*

we have $\|\phi(\mathbf{x})\|_{\widehat{X}} = \|\mathbf{x}\|$ (we call such a map an isometric embedding) such that $\phi(X)$ is dense in $\widehat{X}$. Moreover, this embedding is unique, in the sense that if $\widetilde{X}$ is another Banach space with an isometric embedding $\psi : X \rightarrow \widetilde{X}$ such that $\psi(X)$ is dense in $\widetilde{X}$, then there exists a unique isomorphism of Banach spaces $g : \widetilde{X} \rightarrow \widehat{X}$ such that $g \circ \psi = \phi$.

Remark 9.1.3. This theorem also holds for general metric spaces, that is, every metric space can be embedded as a dense subset in a complete metric space; but we need the additional linear structure in all of our applications, and our proofs are simplified by assuming X is a normed vector space.

Proof. Let X' be the set of all Cauchy sequences in X . The space X maps injectively into X' by sending $\mathbf{x} \in X$ to the constant sequence $(\mathbf{x})_{k=0}^\infty$. For any $\alpha, \beta \in \mathbb{F}$ and any two Cauchy sequences $(\mathbf{x}_k)_{k=0}^\infty$ and $(\mathbf{y}_k)_{k=0}^\infty$, let $\alpha(\mathbf{x}_k)_{k=0}^\infty + \beta(\mathbf{y}_k)_{k=0}^\infty$ be defined to be the sequence $(\alpha\mathbf{x}_k + \beta\mathbf{y}_k)_{k=0}^\infty$. It is straightforward to check that this is again a Cauchy sequence, that X' is a vector space, and that X is a vector subspace.

For each Cauchy sequence $(\mathbf{x}_k)_{k=0}^\infty$, the sequence of norms $(\|\mathbf{x}_k\|)_{k=0}^\infty$ is a Cauchy sequence in $\mathbb{R}$, since for any $\varepsilon > 0$ there is an N such that

$$|\|\mathbf{x}_n\| - \|\mathbf{x}_m\|| \leq \|\mathbf{x}_n - \mathbf{x}_m\| < \varepsilon$$

whenever $n, m > N$. Thus, $(\|\mathbf{x}_k\|)_{k=0}^\infty$ has a limit. Define $\|(\mathbf{x}_k)_{k=0}^\infty\|_{X'} = \lim_{k \rightarrow \infty} \|\mathbf{x}_k\|$. Again, it is straightforward to check that this is a seminorm, but there are many sequences $(\mathbf{x}_k)_{k=0}^\infty$ such that $\|(\mathbf{x}_k)_{k=0}^\infty\| = 0$ but $(\mathbf{x}_k)_{k=0}^\infty \neq (0)_{k=1}^\infty$ (the zero element of X'), so it is not a norm; see Exercise 9.2.

Let $K \subset X'$ be the set of Cauchy sequences $(\mathbf{x}_k)_{k=0}^\infty$ such that $\|(\mathbf{x}_k)_{k=0}^\infty\|_{X'} = 0$. Let $\widehat{X} = X'/K$ be the quotient space. By Proposition 9.1.1, the seminorm $\|\cdot\|_{X'}$ induces a norm on the quotient $\widehat{X}$. The map $\phi : X \rightarrow \widehat{X}$ given by sending any $\mathbf{x} \in X$ to the equivalence class of the constant sequence $(\mathbf{x})_{k=0}^\infty$ is an isometry, because $\|\phi(\mathbf{x})\| = \lim_{k \rightarrow \infty} \|\mathbf{x}\| = \|\mathbf{x}\|$. Also, ϕ is injective, because $\phi(\mathbf{x}) \in K$ if and only if $\|\mathbf{x}\| = \|\phi(\mathbf{x})\| = 0$, which occurs if and only if $\mathbf{x} = 0$.

To see that $\phi(X)$ is dense in $\widehat{X}$, consider any $\zeta = (\mathbf{x}_k)_{k=0}^\infty \in \widehat{X}$. For any $\varepsilon > 0$, choose an $N > 0$ such that $\|\mathbf{x}_n - \mathbf{x}_m\| < \varepsilon/2$ for all $n, m > N$. We have $\|\phi(\mathbf{x}_n) - \zeta\| = \lim_{k \rightarrow \infty} \|\mathbf{x}_n - \mathbf{x}_k\| \leq \varepsilon/2$, so $\phi(\mathbf{x}_n) \in B(\zeta, \varepsilon)$, hence $\zeta \in \overline{\phi(X)}$. Now by Lemma 5.4.26, since every Cauchy sequence in a dense subspace converges in $\widehat{X}$, the space $\widehat{X}$ is complete.

All that remains is the proof of uniqueness. If $\psi : X \rightarrow \widetilde{X}$ is as in the hypothesis of the theorem, we must construct an isomorphism of Banach spaces $g : \widetilde{X} \rightarrow \widehat{X}$, by which we mean an isometry that has an inverse that is also an isometry.

Since $\psi(X)$ is dense in $\widetilde{X}$, for each $\tilde{\mathbf{x}} \in \widetilde{X}$ there is some sequence $(\psi(\mathbf{x}_n))_{n=0}^\infty$ converging to $\tilde{\mathbf{x}}$. Define $g(\tilde{\mathbf{x}}) = \lim_{n \rightarrow \infty} \phi(\mathbf{x}_n)$. To see that this map is well defined, consider any other sequence $(\psi(\mathbf{y}_n))_{n=0}^\infty$ in $\psi(X)$ converging to $\tilde{\mathbf{x}}$. Both $(\psi(\mathbf{x}_n))_{n=0}^\infty$ and $(\psi(\mathbf{y}_n))_{n=0}^\infty$ are Cauchy sequences with $\|\psi(\mathbf{x}_n - \mathbf{y}_n)\| \rightarrow 0$. Since $\|\psi(\mathbf{x})\| = \|\mathbf{x}\| = \|\phi(\mathbf{x})\|$ for all $\mathbf{x} \in X$, we have $\|\phi(\mathbf{x}_n - \mathbf{y}_n)\| \rightarrow 0$, so $\lim_{n \rightarrow \infty} \phi(\mathbf{x}_n) = \lim_{n \rightarrow \infty} \phi(\mathbf{y}_n)$. Thus, $g(\tilde{\mathbf{x}})$ does not depend on the choice of the sequence $(\mathbf{x}_n)_{n=0}^\infty$.

From the definition and the properties of ϕ and ψ , it is straightforward to verify that g is linear and $\|g(\tilde{x})\|_{\hat{X}} = \|\tilde{x}\|_{\tilde{X}}$. An identical argument with the roles of $\tilde{X}$ and $\hat{X}$ switched gives another isometry, $h : \hat{X} \rightarrow \tilde{X}$, which is easily seen to be the inverse of g , and hence g is an isomorphism of Banach spaces. $\square$

Vista 9.1.4. A similar argument can be used to construct the Banach space $\mathbb{R}$ as the completion of the metric space $\mathbb{Q}$. Beware, however, that this method of completing $\mathbb{Q}$ does not fit into the hypotheses of Theorem 9.1.2, because $\mathbb{Q}$ is not a vector space over $\mathbb{R}$ or $\mathbb{C}$. The proof of the theorem above also uses the fact that $\mathbb{R}$ is complete in order to construct the seminorm on $\hat{X}$, so to complete $\mathbb{Q}$ in this way needs some additional steps.

9.1.3 The Space $L^1([a, b]; X)$

The main application of Theorem 9.1.2 is to complete $S([a, b]; X)$ with respect to the L^1 -norm. Recall that any sequence that converges with respect to the L^∞ -norm must also converge in the L^1 -norm (see Corollary 8.2.9). This means that the completion of $S([a, b]; X)$ with respect to the L^1 -norm contains $\mathcal{R}([a, b]; X)$. Therefore, to construct the desired space, we can either take Cauchy sequences in $\mathcal{R}([a, b]; X)$ or Cauchy sequences in $S([a, b]; X)$. But, since step functions are so much simpler to work with, we use $S([a, b]; X)$.

Definition 9.1.5. Let $L^1([a, b]; X)$ denote the completion $\hat{S}([a, b]; X)$ of S with respect to the L^1 -norm. By the continuous linear extension theorem, there is a unique linear extension of the step-function integral $\mathcal{I}$ to all of $L^1([a, b]; X)$. We usually denote this extension by the symbol $\int_{[a, b]}$. If $[a, b] \subset \mathbb{R}$, we often write $\int_a^b f$ instead of $\int_{[a, b]} f$.

Remark 9.1.6. As discussed in the previous chapter, the integral is a bounded transformation, and hence continuous, so it commutes with limits in the L^1 -norm (that is, limits pass through integrals). In particular, this means that if $(s_k)_{k=0}^\infty$ is any L^1 -Cauchy sequence representing an element $f \in L^1([a, b]; X)$, we have

$$\lim_{k \rightarrow \infty} \int_{[a, b]} s_k = \int_{[a, b]} f.$$

Moreover, we have defined the norm on $L^1([a, b]; X)$ to be

$$\|f\|_{L^1} = \lim_{k \rightarrow \infty} \|s_k\|_{L^1} = \lim_{k \rightarrow \infty} \int_{[a, b]} \|s_k(t)\|.$$

Thus, the integral and the L^1 -norm both commute with limits of L^1 -Cauchy sequences.

Proposition 9.1.7. If $(s_k)_{k=0}^\infty$ is any L^1 -Cauchy sequence of real-valued step functions on $[a, b]$, then $(s_k^+)_{k=0}^\infty$ and $(s_k^-)_{k=0}^\infty$ are L^1 -Cauchy, as is $(|s_k|)_{k=0}^\infty$.

Proof. The proof is Exercise 9.3 $\square$

9.2 More about Measure Zero

9.2.1 An Alternative Definition of Sets of Measure Zero

In Definition 8.3.1 the sets of measure zero were described in terms of coverings by unions of intervals. This has many benefits pedagogically, but we prefer now to give an alternative definition. The new definition is a little harder to state and may be less intuitive, but it is easier to use when working with the concept of convergence almost everywhere.

Definition 9.2.1. A set $E \subset [a, b] \subset \mathbb{R}^n$ has measure zero if there is an L^1 -Cauchy sequence $(s_k)_{k=0}^\infty \subset S([a, b]; \mathbb{R})$ of step functions such that $\lim_{k \rightarrow \infty} |s_k(\mathbf{y})| = \infty$ for each $\mathbf{y} \in E$.

For the rest of this section, when we say *measure zero*, we mean it in the sense of Definition 9.2.1, unless otherwise specified. Near the end of this section we prove that the two definitions are equivalent. Unless otherwise specified, we work on a fixed compact interval $[a, b] \subset \mathbb{R}^n$, so all functions are assumed to be defined on $[a, b]$ and all integration is done on $[a, b]$.

9.2.2 Convergence Almost Everywhere of L^1 -Cauchy Sequences

Lemma 9.2.2. If $(s_k)_{k=0}^\infty$ is an L^1 -Cauchy, monotone increasing (or monotone decreasing) sequence of step functions, then there exists a function f such that $s_k \rightarrow f$ a.e.

Proof. Given a monotone increasing L^1 -Cauchy sequence $(s_k)_{k=0}^\infty$ of step functions, define f by

$$f(t) = \begin{cases} \lim_{k \rightarrow \infty} s_k(t) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases} \quad (9.1)$$

For each t , the sequence $(s_k(t))_{k=0}^\infty$ is nondecreasing, so the only way it can fail to converge is if $s_k(t) \rightarrow \infty$. Let E be the set of points t where $s_k(t)$ diverges. Using $(s_k)_{k=0}^\infty$ itself as the L^1 -Cauchy sequence in Definition 9.2.1 shows that E has measure zero, and hence $s_k \rightarrow f$ a.e.

The case where $(s_k)_{k=0}^\infty$ is monotone decreasing follows from a similar argument; see Exercise 9.7. $\square$

Lemma 9.2.3. If $(s_k)_{k=0}^\infty$ is a sequence of step functions on $[a, b]$ such that $\sum_{k=0}^\infty \|s_k\|_{L^1}$ converges to a finite value, then there exists a function F such that $\sum_{k=0}^\infty s_k = F$ a.e.

Proof. For each $n \in \mathbb{N}$ let F_n be the partial sum $F_n = \sum_{k=0}^n s_k$, and let $T_n = \sum_{k=0}^n |s_k|$. Each T_n is a nonnegative step function, and the sequence $(T_n)_{n=0}^\infty$ is monotone increasing. For each n we have

$$\|T_n\|_{L^1} = \left\| \sum_{k=0}^n |s_k| \right\|_{L^1} = \sum_{k=0}^n \|s_k\|_{L^1}.$$

Thus, the monotone increasing sequence $(\|T_n\|_{L^1})_{n=0}^\infty$ converges to some value M . Given $\varepsilon > 0$, choose $N > 0$ such that $0 < (M - \|T_n\|_{L^1}) < \varepsilon$ for all $n > N$. We have

$$\begin{aligned} \|T_n - T_m\|_{L^1} &= \int_{[a,b]} |T_n - T_m| \\ &= \int_{[a,b]} (T_n - T_m) \\ &= \|T_n\|_{L^1} - \|T_m\|_{L^1} \\ &= (M - \|T_m\|_{L^1}) - (M - \|T_n\|_{L^1}) \\ &< \varepsilon \end{aligned}$$

whenever $N < m < n$. Therefore, $(T_n)_{n=0}^\infty$ is L^1 -Cauchy. By Lemma 9.2.2 it converges a.e., and thus $(F_n)_{n=0}^\infty$ converges a.e., since pointwise absolute convergence implies pointwise convergence. The desired function F is given by setting $F(t) = \sum_{k=0}^\infty s_k(t)$ when the sum converges and $F(t) = 0$ when it does not converge. $\square$

Proposition 9.2.4. *If $(s_k)_{k=0}^\infty$ is an L^1 -Cauchy sequence of step functions on $[a, b]$, then there exists a subsequence $(s_{k_\ell})_{\ell=0}^\infty$ and a function f such that $s_{k_\ell} \rightarrow f$ a.e. Moreover, there exist two monotone increasing L^1 -Cauchy sequences $(\phi_\ell)_{\ell=0}^\infty$ and $(\psi_\ell)_{\ell=0}^\infty$ of step functions such that for every $\ell \in \mathbb{N}$, we have $s_{k_\ell} = \phi_\ell - \psi_\ell$.*

Proof. For each $\ell \in \mathbb{N}$ let k_ℓ be chosen such that $\|s_n - s_m\|_{L^1} < 2^{-\ell}$ for all $m, n \geq k_\ell$. Let $g_0 = s_{k_0}$ and for each integer $\ell > 0$ let $g_\ell = s_{k_\ell} - s_{k_{\ell-1}}$. This gives $\|g_\ell\|_{L^1} < 2^{1-\ell}$ for all $\ell > 0$. We have

$$\sum_{\ell=0}^\infty \|g_k\|_{L^1} \leq \|s_{k_0}\|_{L^1} + \sum_{\ell=1}^\infty 2^{1-\ell} = \|s_{k_0}\|_{L^1} + 2.$$

By Lemma 9.2.3 the sequence of partial sums $s_{k_N} = \sum_{\ell=0}^N g_k$ converges to a function f almost everywhere.

Let $\phi_m = \sum_{\ell=0}^m g_\ell^+$ and $\psi_m = \sum_{\ell=0}^m g_\ell^-$, so $s_{k_m} = \phi_m - \psi_m$. For each $\ell \in \mathbb{N}$ we have $g_\ell^+ \geq 0$ and $g_\ell^- \geq 0$, so the sequences $(\phi_m)_{m=0}^\infty$ and $(\psi_m)_{m=0}^\infty$ are both monotone increasing. They are also L^1 -Cauchy because for any $\varepsilon > 0$

$$\begin{aligned} \|\phi_m - \phi_n\|_{L^1} &= \sum_{\ell=n+1}^m \|g_\ell^+\|_{L^1} \leq \sum_{\ell=n+1}^m \|g_\ell^+ + g_\ell^-\|_{L^1} \\ &= \sum_{\ell=n+1}^m \|g_\ell\|_{L^1} < \sum_{\ell=n+1}^m 2^{1-\ell} < 2^{1-n} < \varepsilon \end{aligned}$$

whenever $m > n > 1 - \log_2 \varepsilon$ (and similarly for $(\psi_\ell)_{\ell=0}^\infty$). $\square$

9.2.3 Equivalence of Definitions of Sets of Measure Zero

We now use the previous results about convergence of L^1 -Cauchy sequences to prove that the two definitions of measure zero (Definitions 9.2.1 and 8.3.1) are equivalent.

Proposition 9.2.5. *For a set $E \subset \mathbb{R}^n$ the following are equivalent:*

- (i) *E has measure zero in the sense of Definition 9.2.1.*
- (ii) *There exists a monotone increasing L^1 -Cauchy sequence $(\phi_k)_{k=0}^\infty$ of step functions such that $\phi_k(t) \rightarrow \infty$ for every $t \in E$.*
- (iii) *E has measure zero in the sense of Definition 8.3.1.*

Proof.

(i) $\implies$ (ii) Let $(s_k)_{k=0}^\infty$ be an L^1 -Cauchy sequence of step functions with $s_k(t) \rightarrow \infty$ for every $t \in E$. By Proposition 9.2.4 there is a pair of monotone increasing, L^1 -Cauchy sequences $(\phi_\ell)_{\ell=1}^\infty$ and $(\psi_\ell)_{\ell=1}^\infty$ such that $(\phi_\ell - \psi_\ell)_{\ell=1}^\infty$ is a subsequence of $(s_k)_{k=0}^\infty$. Since $s_k(t) \rightarrow \infty$ for every $t \in E$, we must also have $\phi_k(t) \rightarrow \infty$ for every $t \in E$.

(ii) $\implies$ (iii) Given $(\phi_k)_{k=0}^\infty$ as in (ii), let $F = \{t \in \mathbb{R}^n \mid \lim_{\ell \rightarrow \infty} \phi_\ell(t) = \infty\}$. Since $(\phi_\ell)_{\ell=0}^\infty$ is L^1 -Cauchy, we have $\lim_{\ell \rightarrow \infty} \int_{\mathbb{R}^n} \phi_\ell = M < \infty$. Monotonicity of the sequence implies that $\int_{\mathbb{R}^n} \phi_\ell \leq M$ for all ℓ .

Given $\varepsilon > 0$, let $F_\ell = \{t \in \mathbb{R}^n \mid \phi_\ell(t) > M/\varepsilon\}$ for each $\ell \in \mathbb{N}$. We have $E \subset F \subset \bigcup_{\ell=1}^\infty F_\ell$, and since $F_\ell \subset F_{\ell+1}$ for every ℓ , it suffices to show that each F_ℓ is a finite union of intervals, the sum of whose measure is less than ε . Since ϕ_ℓ is a step function, F_ℓ is a finite union of (partially open or closed) intervals, and $\lambda(F_\ell)M\varepsilon < \int_{\mathbb{R}^n} \phi_\ell \leq M$, so $\lambda(F_\ell) < \varepsilon$.

(iii) $\implies$ (i) For each $m \in \mathbb{N}$ choose intervals $(I_{m,k})_{k=0}^\infty$, satisfying the hypothesis of (iii) for $\varepsilon = 2^{-m}$. For each $k \in \mathbb{N}$ let $s_k = \sum_{m \leq k} \sum_{\ell \leq m} \mathbb{1}_{I_{m,\ell}}$. These may not quite be step functions, because the intervals may not have the right form, but they are *generalized step functions* as defined in Exercise 9.6, and it is proved in Exercise 9.6 that it suffices to use generalized step functions in Definition 9.2.1.

We have

$$\|s_k\|_{L^1} = \sum_{m \leq k} \sum_{\ell \leq m} \lambda(I_{m,\ell}) < \sum_{m \leq k} 2^{-m} < 1.$$

Moreover, the sequence $(s_k)_{k=0}^\infty$ is monotone increasing, and hence (by essentially the same argument as given in the proof of Lemma 9.2.3) the sequence $(s_k)_{k=0}^\infty$ is L^1 -Cauchy. But for any $t \in E$, we have

$$s_k(t) = \sum_{m \leq k} \sum_{\ell \leq m} 1 \rightarrow \infty. \quad \square$$

Proposition 9.2.6. *If $E \subset \mathbb{R}^n$ has measure zero in $\mathbb{R}^n$, then $E \times \mathbb{R}^m \subset \mathbb{R}^{n+m}$ has measure zero in $\mathbb{R}^{n+m}$.*

Proof. The proof is Exercise 9.8. $\square$

Corollary 9.2.7. *The boundary of any interval (and hence the set of discontinuous points of any step function) has measure zero.*

Proof. The boundary of any interval in $\mathbb{R}^n$ is contained in a finite union of sets of the form $\{a\} \times [c, d]$, where $[c, d] \subset \mathbb{R}^{n-1}$, and where $a \in \mathbb{R}$ is a single point (hence of measure zero). $\square$

9.3 Lebesgue-Integrable Functions

We have defined $L^1([a, b]; \mathbb{R})$ to be a set of equivalence classes of L^1 -Cauchy sequences. But equivalence classes of sequences are awkward to work with, and it would be much nicer if we could just use functions. In this section we show how to do this by proving that for each equivalence class of L^1 -Cauchy sequences in $L^1([a, b]; \mathbb{R})$ there is a uniquely determined equivalence class of functions (with respect to the equivalence relation $=$ a.e.).

If we define the set $\mathcal{L}$ of integrable functions to be $\mathcal{L} = \{f \mid \exists (s_k)_{k=0}^\infty \in L^1([a, b]; \mathbb{R}), s_k \rightarrow f \text{ a.e.}\}$, and if we let $\mathcal{L}_0 = \{f \in \mathcal{L} \mid f = 0 \text{ a.e.}\}$, then this gives a linear isomorphism of vector spaces $\Phi : L^1([a, b]; \mathbb{R}) \rightarrow \mathcal{L}/\mathcal{L}_0$. And if we define a norm on $\mathcal{L}/\mathcal{L}_0$ by $\|f\|_{L^1} = \lim_{k \rightarrow \infty} \|s_k\|$ whenever $s_k \rightarrow f$ a.e., then $\|f\|_{L^1}$ is well defined, and Φ preserves the norm. That is to say, $L^1([a, b]; \mathbb{R})$ is isomorphic, as a normed linear space, to the space $\mathcal{L}/\mathcal{L}_0$ of equivalence classes of functions that are pointwise limits of some L^1 -Cauchy sequence of step functions.

The upshot of all of this is that we may talk about elements of $L^1([a, b]; \mathbb{R})$ as integrable functions, rather than as Cauchy sequences of step functions.

9.3.1 L^1 -Cauchy Sequences and Integrable Functions

Definition 9.3.1. *Fix a compact n -interval $[a, b] \subset \mathbb{R}^n$. We say that a function $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue integrable or just integrable on $[a, b]$ if there exists an L^1 -Cauchy sequence of step functions $(s_k)_{k=0}^\infty$ such that $s_k \rightarrow f$ a.e. We denote the vector space of integrable functions on $[a, b]$ by $\mathcal{L}$ (for this section only).*

The results of Section 9.2.2 show that for any L^1 -Cauchy sequence $(s_k)_{k=0}^\infty$ of step functions, we can construct a function f such that some subsequence converges almost everywhere to f . We use this to define a map $\Phi : L^1([a, b]; \mathbb{R}) \rightarrow \mathcal{L}/\mathcal{L}_0$ by sending any L^1 -Cauchy sequence $(s_k)_{k=0}^\infty$ of step functions to the integrable function f guaranteed to exist by Proposition 9.2.4. But it is not yet clear that this map is well defined. To see this, we must show that the equivalence class $f + \mathcal{L}_0$ is uniquely determined.

Proposition 9.3.2. *If $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ are L^1 -Cauchy sequences of step functions that are equivalent in $L^1([a, b]; \mathbb{R})$, and if $(s_k)_{k=0}^\infty \rightarrow f$ a.e. and $(s'_k)_{k=0}^\infty \rightarrow g$ a.e., then $f = g$ a.e.*

Proof. For each $k \in \mathbb{N}$, the difference $u_k = s_k - s'_k$ is a step function, and by the definition of equivalence in $L^1([a, b]; \mathbb{R})$, the sequence $(u_k)_{k=0}^\infty$ is an L^1 -Cauchy sequence with $\|(u_k)_{k=0}^\infty\|_{L^1} \rightarrow 0$. Moreover, $(u_k)_{k=0}^\infty$ converges almost everywhere to $h = f - g$, so it suffices to show that the set of points t where $h(t) \neq 0$ has measure zero.

Let $E_m = \{t \in \mathbb{R}^n \mid |h(t)| > 1/m\}$. Since the set where $h(x) \neq 0$ is precisely the (countable) union $\bigcup_{m=1}^\infty E_m$, it suffices to show that each E_m has measure zero.

Each u_k can be written as $u_k = \sum_{j \in J_k} c_{k,j} \mathbb{1}_{I_{k,j}}$ where the index j runs over a finite set J_k , each $I_{k,j}$ is an interval, and each $c_{k,j}$ is a real number. For each $m \in \mathbb{N}$ and for each $\varepsilon > 0$ choose u_k such that $\|u_k\|_{L^1} < \varepsilon/m$. The finite set of intervals $\mathcal{J} = \{I_{k,j} \mid j \in J_k, |c_{k,j}| > 1/m\}$ covers E_m , and we have

$$\sum_{\mathcal{J}} \lambda(I_{k,j}) = m \sum_{j \in J_k} \frac{1}{m} \lambda(I_{k,j}) \leq m \sum_{j \in J_k} |c_{k,j}| \lambda(I_{k,j}) = m \|u_k\|_{L^1} < \varepsilon.$$

Therefore, E_m has measure zero by Proposition 9.2.5. $\square$

This proposition shows that the function $\Phi : L^1([a, b]; \mathbb{R}) \rightarrow \mathcal{L}/\mathcal{L}_0$ is well defined, because if $(s_k)_{k=0}^\infty$ is any L^1 -Cauchy sequence with a subsequence $(s_{k_\ell})_{\ell=1}^\infty$ converging almost everywhere to f (so that we should have $\Phi((s_k)) = f$), and if $(s'_k)_{k=0}^\infty$ is an equivalent L^1 -Cauchy sequence with a subsequence $(s'_{j_m})_{m=1}^\infty$ converging almost everywhere to g (so that we should have $\Phi((s'_k)) = g$), then the subsequences are equivalent in $L^1([a, b]; \mathbb{R})$, so $f = g$ a.e.

The map Φ is surjective by the definition of $\mathcal{L}$, and it is straightforward to check that Φ is also linear. We now show that Φ is injective. To do this, we must show that if $f = g$ a.e., and if $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ are two L^1 -Cauchy sequences of step functions such that $s_k \rightarrow f$ a.e. and $s'_k \rightarrow g$ a.e., then $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ are equivalent in $L^1([a, b]; \mathbb{R})$. To do this we first need two lemmata.

Lemma 9.3.3. *If $(s_k)_{k=0}^\infty$ is a monotone decreasing sequence of nonnegative step functions such that $s_k \rightarrow 0$ a.e., then $\|s_k\|_{L^1} \rightarrow 0$. And similarly, if $(s_k)_{k=0}^\infty$ is a monotone increasing sequence of nonpositive step functions such that $s_k \rightarrow 0$ a.e., then $\|s_k\|_{L^1} \rightarrow 0$.*

Proof. Let $(s_k)_{k=0}^\infty$ be a monotone decreasing sequence of nonnegative step functions such that $s_k \rightarrow 0$ a.e. Let D_0 be the set of all t where $(s_k(t))_{k=0}^\infty$ does not converge to 0, and for each $m \in \mathbb{N}$, let D_m be the set of all t such that s_m is discontinuous at t . Each D_i has measure zero, so the countable union $D = \bigcup_{i=0}^\infty D_i$ has measure zero.

The step function s_1 is bounded by some positive real number B and has a compact interval $[a, b]$ as its domain. Since the sequence is monotone decreasing, each function s_k is also bounded by B , and we may assume that s_k has the same domain $[a, b]$. For any $\varepsilon > 0$ choose open intervals $\{I_k\}_{k \in \mathbb{N}}$ such that $D \subset \bigcup_{k \in \mathbb{N}} I_k$ and such that $\sum_{k \in \mathbb{N}} \lambda(I_k) < \delta = \varepsilon/(2B)$.

For each $t \in [a, b] \setminus D$, there exists an $m_t \in \mathbb{N}$ such that $s_{m_t}(t) < \gamma = \varepsilon/(2\lambda([a, b]))$. Since t is not a point of discontinuity of s_{m_t} , there is an open neighborhood U_t of t where $s_{m_t}(t') < \gamma$ for all $t' \in U_t$. The collection $\{I_k\}_{k \in \mathbb{N}} \cup \{U_t \mid t \in [a, b] \setminus D\}$ is an open cover of the compact interval $[a, b]$, so there is a finite subcover $I_{i_1}, \dots, I_{i_N}, U_{t_1}, \dots, U_{t_j}$ that covers $[a, b]$. Let $M = \max(m_{t_1}, \dots, m_{t_j})$.

Since $(s_k)_{k=0}^\infty$ is monotone decreasing, if $s_{m_t}(t') < \gamma$, then $s_\ell(t') < \gamma$ for every $\ell \geq M$. Therefore, we have $0 \leq s_\ell(t') < \gamma$ for all t' in $\bigcup_{i=1}^j U_{t_i}$.

Putting this all together gives

$$\|s_\ell\|_{L^1} = \int_{[a,b]} s_\ell < B \sum_{k=1}^N \lambda(I_k) + \gamma \lambda([a,b]) = \varepsilon/2 + \varepsilon/2 = \varepsilon$$

for all $\ell > M$. This shows that $\|s_k\|_{L^1} \rightarrow 0$.

The result about a monotone increasing sequence $(s_k)_{k=0}^\infty$ of nonpositive step functions follows from the nonnegative decreasing case by changing the signs. $\square$

Lemma 9.3.4. *Let $(\phi_k)_{k=0}^\infty$ and $(\psi_k)_{k=0}^\infty$ be monotone increasing sequences of step functions such that $\phi_k \rightarrow f$ a.e. and $\psi_k \rightarrow g$ a.e. If $f \leq g$ a.e., then*

$$\lim_{k \rightarrow \infty} \int_{[a,b]} \phi_k \leq \lim_{k \rightarrow \infty} \int_{[a,b]} \psi_k.$$

Proof. For each $m \in \mathbb{N}$ the sequence $(\phi_m - \psi_k)_{k=0}^\infty$ is monotone decreasing, and

$$\lim_{k \rightarrow \infty} \phi_m - \psi_k = \phi_m - \lim_{k \rightarrow \infty} \psi_k = \phi_m - g \leq f - g \leq 0 \text{ a.e.}$$

Thus, the sequence $((\phi_m - \psi_k)^+)_{k=0}^\infty$ is nonnegative and monotone decreasing and must converge to zero almost everywhere.

By Lemma 9.3.3 we have

$$\lim_{k \rightarrow \infty} \int_{[a,b]} (\phi_m - \psi_k)^+ = 0.$$

Moreover, we have $(\phi_m - \psi_k) \leq (\phi_m - \psi_k)^+$, so

$$\int_{[a,b]} (\phi_m - \psi_k) \leq \int_{[a,b]} (\phi_m - \psi_k)^+,$$

which gives

$$\int_{[a,b]} \phi_m \leq \lim_{k \rightarrow \infty} \left(\int_{[a,b]} (\phi_m - \psi_k)^+ + \int_{[a,b]} \psi_k \right) = \lim_{k \rightarrow \infty} \int_{[a,b]} \psi_k.$$

Now letting $m \rightarrow \infty$ gives the desired result. $\square$

Theorem 9.3.5. *Let $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ be L^1 -Cauchy sequences of step functions (not necessarily monotone) such that $s_k \rightarrow f$ a.e. and $s'_k \rightarrow g$ a.e. If $f \leq g$ a.e., then*

$$\lim_{k \rightarrow \infty} \int_{[a,b]} s_k \leq \lim_{k \rightarrow \infty} \int_{[a,b]} s'_k.$$

Proof. By Proposition 9.2.4 we may choose monotone increasing L^1 -Cauchy sequences $(\phi_k)_{k=0}^\infty$, $(\psi_k)_{k=0}^\infty$, $(\alpha_k)_{k=0}^\infty$, and $(\beta_k)_{k=0}^\infty$ such that $(\phi_k - \psi_k)_{k=0}^\infty$ is a subsequence of $(s_k)_{k=0}^\infty$ and $(\alpha_k - \beta_k)_{k=0}^\infty$ is a subsequence of $(s'_k)_{k=0}^\infty$.

Any subsequence of an L^1 -Cauchy sequence is equivalent in $L^1([a, b]; \mathbb{R})$ to the original sequence, so $(\phi_k - \psi_k) \rightarrow f$ a.e. and $(\alpha_k - \beta_k) \rightarrow g$ a.e. Since each sequence $(\phi_k)_{k=0}^\infty$, $(\psi_k)_{k=0}^\infty$, $(\alpha_k)_{k=0}^\infty$, and $(\beta_k)_{k=0}^\infty$ is L^1 -Cauchy, there exist functions ϕ , ψ , α , and β such that each sequence converges to the corresponding function almost everywhere. We have $\phi - \psi = f \leq g = \alpha - \beta$ a.e., so

$$\phi + \beta \leq \alpha + \psi \text{ a.e.}$$

The sequences $(\phi_k + \beta_k)_{k=0}^\infty$ and $(\alpha_k + \psi_k)_{k=0}^\infty$ are monotone increasing, so by Lemma 9.3.4 we have

$$\lim_{k \rightarrow \infty} \int_{[a, b]} (\phi_k + \beta_k) \leq \lim_{k \rightarrow \infty} \int_{[a, b]} (\alpha_k + \psi_k).$$

Since integration on step functions is linear, and addition (and subtraction) are continuous, we have

$$\lim_{k \rightarrow \infty} \int_{[a, b]} (\phi_k - \psi_k) \leq \lim_{k \rightarrow \infty} \int_{[a, b]} (\alpha_k - \beta_k).$$

By the uniqueness in the continuous linear extension theorem, limits of integrals of L^1 -Cauchy sequences are independent of the representative of the equivalence class, so we have

$$\lim_{k \rightarrow \infty} \int_{[a, b]} s_k = \lim_{k \rightarrow \infty} \int_{[a, b]} (\phi_k - \psi_k) \leq \lim_{k \rightarrow \infty} \int_{[a, b]} (\alpha_k - \beta_k) = \lim_{k \rightarrow \infty} \int_{[a, b]} s'_k. \quad \square$$

Proposition 9.3.6. *If $f = g$ a.e., and if $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ are two L^1 -Cauchy sequences of step functions such that $s_k \rightarrow f$ a.e. and $s'_k \rightarrow g$ a.e., then $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ are equivalent in $L^1([a, b]; \mathbb{R})$.*

Proof. Proposition 9.1.7 shows that the sequences $(|s_k|)_{k=0}^\infty$ and $(|s'_k|)_{k=0}^\infty$ are L^1 -Cauchy, with $|s_k| \rightarrow |f|$ a.e. and $|s'_k| \rightarrow |g|$ a.e. By Theorem 9.3.5 we have

$$\lim_{k \rightarrow \infty} \|s_k\|_{L^1} = \lim_{k \rightarrow \infty} \int_{[a, b]} |s_k| = \lim_{k \rightarrow \infty} \int_{[a, b]} |s'_k| = \lim_{k \rightarrow \infty} \|s'_k\|_{L^1}. \quad (9.2)$$

Substituting $(s_k - s'_k)$ for s_k and 0 for s'_k in (9.2) gives the desired result. $\square$

We have shown that Φ is an isomorphism of vector spaces. We may use Φ to define the L^1 -norm on $\mathcal{L}/\mathcal{L}_0$ by $\|f\|_{L^1} = \|(s_k)_{k=0}^\infty\|_{L^1} = \lim_{k \rightarrow \infty} \|s_k\|_{L^1}$ whenever $f = \Phi((s_k)_{k=0}^\infty)$. Similarly, we may define the integral $\int_{[a, b]} f = \lim_{k \rightarrow \infty} \int_{[a, b]} s_k$ whenever $f = \Phi((s_k)_{k=0}^\infty)$.

From now on we usually describe elements of $L^1([a, b]; \mathbb{R})$ as integrable functions (or equivalence classes of integrable functions) rather than as equivalence classes of L^1 -Cauchy sequences of step functions—we have showed that the two formulations are equivalent, but functions are usually more natural to work with.

9.3.2 Basic Integral Properties

In this section we prove several properties of integrals and integrable functions. We first prove Proposition 8.4.2, which we restate here for the reader's convenience.

Proposition 8.4.2. *For any $f, g \in L^1([a, b]; \mathbb{R})$ we have the following:*

- (i) *If $f \leq g$ a.e., then $\int_{[a,b]} f \leq \int_{[a,b]} g$.*
- (ii) *$\left| \int_{[a,b]} f \right| \leq \int_{[a,b]} |f| = \|f\|_{L^1}$.*
- (iii) *The functions $\max(f, g)$, $\min(f, g)$, f^+ , f^- , and $|f|$ are all integrable.*
- (iv) *If $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a measurable function (see Definition 8.3.9), and if $|h| \in L^1([a, b]; \mathbb{R})$, then $h \in L^1([a, b]; \mathbb{R})$.*
- (v) *If $\|g\|_{L^\infty} \leq M < \infty$, then $fg \in L^1([a, b]; \mathbb{R})$ and $\|fg\|_{L^1} \leq M\|f\|_{L^1}$.*

Proof. Property (i) is an immediate consequence of Theorem 9.3.5. Property (ii) follows from (i), since $f \leq |f|$ and $-f \leq |f|$. Property (iii) is Exercise 9.12.

For (iv), since h is measurable, there is a sequence $(s_k)_{k=0}^\infty$ of step functions with $s_k \rightarrow h$ a.e. For each $k \in \mathbb{N}$ let

$$\phi_k = \min(-|h|, s_k, |h|) = \max(-|h|, \min(s_k, |h|)) = \begin{cases} |h| & \text{if } s_k \geq |h|, \\ s_k & \text{if } -|h| \leq s_k \leq |h|, \\ -|h| & \text{if } s_k \leq -|h|. \end{cases}$$

By Proposition 8.4.2(iii), each ϕ_k is integrable. Since $s_k \rightarrow h$ a.e., we also have $\phi_k \rightarrow h$ a.e. Since $\phi_k \leq |h|$ for all k , the dominated convergence theorem guarantees that $h \in L^1([a, b]; \mathbb{R})$.

For (v) choose sequences of step functions $(s_k)_{k=0}^\infty$ and $(s'_k)_{k=0}^\infty$ such that $s_k \rightarrow f$ a.e. and $s'_k \rightarrow g$ a.e. The sequence $(s_k s'_k)_{k=0}^\infty$ converges to fg almost everywhere. We have

$$\lim_{k \rightarrow \infty} \int_{[a,b]} |s_k s'_k| \leq \lim_{k \rightarrow \infty} \int_{[a,b]} |s_k| M = M \int_{[a,b]} |f| = M \|f\|_{L^1} < \infty,$$

so by Fatou's lemma we have $|fg| \in L^1([a, b]; \mathbb{R})$, and by (iv) we also have $fg \in L^1([a, b]; \mathbb{R})$. $\square$

Proposition 9.3.7. *Let $A \subset \mathbb{R}^n$ be a measurable set. For any $\mathbf{c} \in \mathbb{R}^n$ and any $f \in L^1(A; \mathbb{R})$, let $f_{\mathbf{c}}$ be the function on $A - \mathbf{c} = \{\mathbf{t} \in \mathbb{R}^n \mid \mathbf{t} + \mathbf{c} \in A\}$ given by $f_{\mathbf{c}}(\mathbf{t}) = f(\mathbf{t} + \mathbf{c})$. We have $f_{\mathbf{c}} \in L^1(A - \mathbf{c}; \mathbb{R})$ and*

$$\int_{A - \mathbf{c}} f_{\mathbf{c}} = \int_A f.$$

Proof. The proof is Exercise 9.15. $\square$

9.4 Proof of Fubini's Theorem

In this section we prove Fubini's theorem, which we restate here for the reader's convenience.

Theorem 8.6.1. *Assume that $f: X \times Y \rightarrow \mathbb{R}$ is integrable on $X \times Y \subset \mathbb{R}^{m+n}$. For each $\mathbf{x} \in X$ consider the function $f_{\mathbf{x}}: Y \rightarrow \mathbb{R}$ given by $f_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$. We have the following:*

- (i) *For almost all $\mathbf{x} \in X$ the function $f_{\mathbf{x}}$ is integrable on Y .*
- (ii) *The function $F: X \rightarrow \mathbb{R}$ given by*

$$F(\mathbf{x}) = \begin{cases} \int_Y f_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} & \text{if } f_{\mathbf{x}} \text{ is integrable,} \\ 0 & \text{otherwise} \end{cases}$$

is integrable on X .

- (iii) *The integral of f may be computed as*

$$\int_{X \times Y} f(\mathbf{x}, \mathbf{y}) d\mathbf{x}d\mathbf{y} = \int_X F(\mathbf{x}) d\mathbf{x}.$$

The first step in the proof of Fubini's theorem is to check that it holds for step functions.

Proposition 9.4.1 (Fubini's Theorem for Step Functions). *If $s: [a, b] \times [c, d] \rightarrow \mathbb{R}$ is a step function, then*

- (i) *for every $\mathbf{x} \in [a, b]$ the function $s_{\mathbf{x}}: [c, d] \rightarrow \mathbb{R}$ is a step function;*
- (ii) *the function $S: [a, b] \rightarrow \mathbb{R}$ given by $S(\mathbf{x}) = \int_{\mathbb{R}^m} s_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}$ is a step function;*
- (iii) *the integral of s can be computed as*

$$\int_{[a,b] \times [c,d]} s(\mathbf{x}, \mathbf{y}) d\mathbf{x}d\mathbf{y} = \int_{[a,b]} S(\mathbf{x}) d\mathbf{x} = \int_{[a,b]} \left(\int_{[c,d]} s_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} \right) d\mathbf{x}.$$

Proof. The proof is Exercise 9.18. $\square$

Lemma 9.4.2. *Let $X = [a, b] \subset \mathbb{R}^n$ and $Y = [c, d] \subset \mathbb{R}^m$. If $E \subset X \times Y$ has measure zero, then for almost all $\mathbf{x} \in X$ the set $E_{\mathbf{x}} = \{\mathbf{y} \in Y \mid (\mathbf{x}, \mathbf{y}) \in E\}$ has measure zero in Y .*

Proof. Using Proposition 9.2.5, choose a monotone increasing L^1 -Cauchy sequence $(\phi_k)_{k=0}^{\infty}$ of step functions on $X \times Y$ such that $\phi_k(\mathbf{x}, \mathbf{y}) \rightarrow \infty$ for all $(\mathbf{x}, \mathbf{y}) \in E$. For

each $k \in \mathbb{N}$ and each $\mathbf{x} \in X$, let $\Phi_k(\mathbf{x}) = \int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y}$, where $\phi_{k,\mathbf{x}}(\mathbf{y}) = \phi_k(\mathbf{x}, \mathbf{y})$. Because of Fubini's theorem for step functions (Proposition 9.4.1) we have

$$\int_X \Phi_k(\mathbf{x}) d\mathbf{x} = \int_X \left(\int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y} \right) d\mathbf{x} = \int_{X \times Y} \phi_k(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

Therefore, the sequence $(\int_X \Phi_k(\mathbf{x}) d\mathbf{x})_{k=0}^\infty$ is bounded, and by the monotone convergence theorem (Theorem 8.4.5) we have $\Phi_k \rightarrow \Phi$ a.e. for some $\Phi \in L^1(X; \mathbb{R})$.

Let $\mathbf{x} \in X$ be such that $\Phi_k(\mathbf{x}) \rightarrow \Phi(\mathbf{x})$, so the sequence $(\int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y})_{k=0}^\infty$ is bounded. Again by the monotone convergence theorem, $\phi_{k,\mathbf{x}}$ converges for almost all $\mathbf{y} \in Y$. On the other hand, if $\mathbf{y} \in E_{\mathbf{x}}$, then $\phi_{k,\mathbf{x}}$ diverges to infinity, so the monotone increasing sequence of step functions $\phi_{k,\mathbf{x}} : Y \rightarrow \mathbb{R}$ shows that $E_{\mathbf{x}}$ has measure zero in Y . $\square$

Proof of Fubini's Theorem. Let $(s_k)_{k=0}^\infty$ be an L^1 -Cauchy sequence of step functions on $X \times Y$ that converges almost everywhere to f . By Proposition 9.2.4 we may assume that each s_k is a difference $s_k = \phi_k - \psi_k$, where $(\phi_k)_{k=0}^\infty$ and $(\psi_k)_{k=0}^\infty$ are monotone increasing L^1 -Cauchy sequences on $X \times Y$. Since integration is linear, it suffices to prove the theorem just in the case that $\lim_{k \rightarrow \infty} \phi_k = f$ a.e.

For each $\mathbf{x} \in X$ we have a sequence $(\phi_{k,\mathbf{x}})_{k=0}^\infty$ of step functions on Y . For each $k \in \mathbb{N}$ define $\Phi_k : X \rightarrow \mathbb{R}$ by $\Phi_k(\mathbf{x}) = \int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y}$, so that $(\Phi_k)_{k=0}^\infty$ is a sequence of step functions on X . Because of Fubini's theorem for step functions (Proposition 9.4.1) we have

$$\begin{aligned} \lim_{k \rightarrow \infty} \int_X \Phi_k(\mathbf{x}) d\mathbf{x} &= \lim_{k \rightarrow \infty} \int_X \left(\int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y} \right) d\mathbf{x} \\ &= \lim_{k \rightarrow \infty} \int_{X \times Y} \phi_k(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= \int_{X \times Y} f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \end{aligned}$$

Therefore, the sequence $(\int_X \Phi_k(\mathbf{x}) d\mathbf{x})_{k=0}^\infty$ is bounded, and by the monotone convergence theorem (Theorem 8.4.5) we have $\Phi_k \rightarrow \Phi$ a.e. for some $\Phi \in L^1(X; \mathbb{R})$ with

$$\int_X \Phi(\mathbf{x}) d\mathbf{x} = \int_{X \times Y} f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

We must now show that $\Phi = F$ a.e. and that $f_{\mathbf{x}}$ is integrable for almost all $\mathbf{x} \in X$. Let E be the measure-zero subset of $X \times Y$ where $(\phi_k(\mathbf{x}, \mathbf{y}))_{k=0}^\infty$ fails to converge. By Lemma 9.4.2 the set $E_{\mathbf{x}} = \{\mathbf{y} \in Y \mid (\mathbf{x}, \mathbf{y}) \in E\}$ has measure zero for almost all $\mathbf{x} \in X$. For any $\mathbf{x} \in X$ such that $E_{\mathbf{x}}$ has measure zero and such that $\Phi_k(\mathbf{x}) \rightarrow \Phi(\mathbf{x})$ we have that $\Phi_k(\mathbf{x}) = \int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y}$ converges, and so by the monotone convergence theorem $f_{\mathbf{x}} = \lim_{k \rightarrow \infty} \phi_{k,\mathbf{x}}$ a.e. is integrable on Y , and

$$\Phi(\mathbf{x}) = \lim_{k \rightarrow \infty} \int_Y \phi_{k,\mathbf{x}}(\mathbf{y}) d\mathbf{y} = \int_Y f_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} = F(\mathbf{x}),$$

as required. $\square$

9.5 Proof of the Change of Variables Theorem

In this section we prove the change of variables theorem. Before we begin the proof, we need to set up a little more notation and develop a few more results about integrable and measurable functions.

Definition 9.5.1. For each $k \in \mathbb{N}$ let $Q_k \subset \mathbb{R}^n$ be the set of points $\mathbf{a} \in \mathbb{R}^n$ whose coordinates a_i are all rational of the form $a_i = c/2^k$ for $c \in \mathbb{Z}$. Let Θ_k be the set of compact intervals (n -cubes) of the form $[\mathbf{a}, \mathbf{a} + (2^{-k}, \dots, 2^{-k})]$, where $\mathbf{a} \in Q_k$.

Proposition 9.5.2. Countable intersections and unions of measurable sets are measurable. If X, Y are measurable, then $X \setminus Y$ is measurable, and every open set (and every closed set) in $\mathbb{R}^n$ is measurable.

Proof. If $\{A_k\}_{k \in \mathbb{N}}$ is a countable collection of measurable sets, then setting $f_n = \prod_{k=1}^n \mathbb{1}_{A_k}$ defines a sequence of measurable functions (by Exercise 9.14(ii)). Moreover, $f_n \rightarrow \mathbb{1}_{\bigcap_{k \in \mathbb{N}} A_k}$ pointwise, so by Exercise 9.14(iii) $\bigcap_{k \in \mathbb{N}} A_k$ is measurable. If X and Y are measurable, then $(\mathbb{1}_X - \mathbb{1}_Y)^+ = \mathbb{1}_{X \setminus Y}$ is measurable. Combining this with the previous result shows that countable unions of measurable functions are measurable.

The empty set is measurable because its indicator function is 0. To see that $\mathbb{R}^n$ is measurable, let f_k be the step function $f_k = \mathbb{1}_{[(-k, \dots, -k), (k, \dots, k)]}$. The sequence $(f_k)_{k=0}^\infty$ converges to $1 = \mathbb{1}_{[\mathbf{a}, \mathbf{b}]}$, so $\mathbb{R}^n$ is measurable.

Let U be an open set. For each $k \in \mathbb{N}$ let C_k be the union of all cubes in Θ_k that lie entirely in U . Since C_k is a countable union of intervals (which are measurable), it is measurable. It is clear that for every $k \in \mathbb{N}$ we have $C_k \subset C_{k+1}$. For every $\mathbf{t} \in U$ there is a $\delta > 0$ such that $B(\mathbf{t}, \delta) \subset U$. For k large enough, we have $2^{-k} < \delta/\sqrt{n}$, which implies that any cube of Θ_k containing $\mathbf{t}$ lies entirely inside $B(\mathbf{t}, \delta) \subset U$, and hence every $\mathbf{t} \in U$ lies in some C_k . Therefore, $U = \bigcup_{k \in \mathbb{N}} C_k$ is measurable. Since every closed set is the complement of an open set, every closed set is also measurable. $\square$

Corollary 9.5.3. Every interval $I \subset \mathbb{R}^n$, whether open, partially open, or closed, is measurable. If $A, B \subset \mathbb{R}^n$ are open with $I \subset B$, and if $f: A \rightarrow B$ is continuous, then $f^{-1}(I)$ is also measurable.

Proof. The proof is Exercise 9.21. $\square$

We are now ready to start the proof of the change of variables formula (Theorem 8.7.5), which we restate here for the reader's convenience.

Theorem 8.7.5. Let U and V be open subsets of $\mathbb{R}^n$, let $X \subset U$ be a measurable subset, and let $\Psi: U \rightarrow V$ be a diffeomorphism. The set $Y = \Psi(X)$ is measurable, and if $f: Y \rightarrow \mathbb{R}$ is integrable, then $(f \circ \Psi) |\det(D\Psi)|$ is integrable on X and

$$\int_Y f = \int_X (f \circ \Psi) |\det(D\Psi)|. \quad (9.3)$$

The fact that $Y = \Psi(X)$ is measurable follows from Corollary 9.5.3. We prove the formula first for the case of a compact interval, and then we extend it to unbounded X .

Lemma 9.5.4. *Given the hypothesis of Theorem 8.7.5, if $n = 1$ and $X = [a, b] \subset U \subset \mathbb{R}$ is a compact interval, then the change of variables theorem holds for all $f \in L^1(Y; \mathbb{R})$, where $Y = \Psi(X)$.*

Proof. If $f \in \mathcal{R}(Y; \mathbb{R})$ (that is, if f is regulated-integrable), then this lemma is just the usual substitution formula (8.11) (or (6.19)). In particular, the formula holds for all step functions.

Since X is a compact interval and Ψ is continuous, the set $Y = \Psi(X)$ is also a compact interval. If $f \in L^1(Y; \mathbb{R})$, then Proposition 9.2.4 implies there are monotone increasing L^1 -Cauchy sequences $(s_k)_{k=0}^\infty$ and $(u_k)_{k=0}^\infty$ of step functions on Y such that $s_k - u_k \rightarrow f$ a.e., and there are functions $g, h \in L^1(Y; \mathbb{R})$ with $f = g - h$ such that $s_k \rightarrow g$ a.e. and $u_k \rightarrow h$ a.e.

By the monotone converge theorem (Theorem 8.4.5) we have

$$\int_Y g = \lim_{k \rightarrow \infty} \int_Y s_k = \lim_{k \rightarrow \infty} \int_X (s_k \circ \Psi) |D\Psi| = \int_X (g \circ \Psi) |D\Psi|,$$

and similarly for h . This gives

$$\int_Y f = \int_Y (g - h) = \int_X ((g \circ \Psi) |D\Psi| - (h \circ \Psi) |D\Psi|) = \int_X (f \circ \Psi) |D\Psi|,$$

as required. $\square$

Lemma 9.5.5. *Given the hypothesis of Theorem 8.7.5, if $X = [a, b] \subset U \subset \mathbb{R}^n$ is a compact interval and $\Psi : U \rightarrow V$ is of the form $\Psi(t) = \Psi(t_1, \dots, t_n) = (t_i, \Psi_2(t), \dots, \Psi_n(t))$ for some i , or of the form $\Psi(t) = (\Psi_1, \dots, \Psi_{n-1}(t), t_i)$, then the change of variables formula (9.3) with respect to Ψ holds for all $f \in L^1(Y; \mathbb{R})$.*

Proof. Here we prove the first case. The second is similar. We proceed by induction on n . The base case of $n = 1$ follows from Lemma 9.5.4.

Given

$$X = [a, b] = [a_1, b_1] \times \cdots \times [a_i, b_i] \times \cdots \times [a_n, b_n],$$

let $S = [a_i, b_i] \subset \mathbb{R}$ and

$$W = [a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [a_{i+1}, b_{i+1}] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^{n-1}.$$

For each $t \in S$ let $\iota_t : W \rightarrow X$ be the injective map

$$\iota_t(\mathbf{w}) = \mathbf{w}_t = (w_1, \dots, w_{i-1}, t, w_{i+1}, \dots, w_n),$$

and let U_t be the open set

$$U_t = \{\mathbf{w} \in \mathbb{R}^{n-1} \mid \iota_t(\mathbf{w}) \in U\}.$$

Also, let $\Psi_t : U_t \rightarrow \mathbb{R}^{n-1}$ be the map

$$\Psi_t(\mathbf{w}) = (\Psi_2(\mathbf{w}_t), \dots, \Psi_n(\mathbf{w}_t))$$

and $V_t = \Psi_t(U_t)$. Since Ψ is a diffeomorphism, it is straightforward to verify that $\Psi_t : U_t \rightarrow V_t$ is bijective. The first row of $D\Psi(\mathbf{w}_t)$ is zero in all but the i th position, and the cofactor expansion of the determinant along this row (Theorem 2.9.16) shows that

$$|\det D\Psi_t(\mathbf{w})| = |\det(D\Psi(\mathbf{w}_t))| \neq 0. \quad (9.4)$$

Therefore, Ψ_t is a diffeomorphism.

Using (9.4) with the Fubini theorem (Theorem 8.6.1), we have

$$\begin{aligned} \int_X (f \circ \Psi) |\det(D\Psi)| &= \int_{S \times W} (f \circ \Psi)(\mathbf{w}_t) |\det(D\Psi)(\mathbf{w}_t)| d\mathbf{w} dt \\ &= \int_S \left(\int_W (f \circ \Psi)(\mathbf{w}_t) |\det(D\Psi)(\mathbf{w}_t)| d\mathbf{w} \right) dt \\ &= \int_S \left(\int_W f(t, \Psi_t(\mathbf{w})) |\det(D\Psi_t)(\mathbf{w})| d\mathbf{w} \right) dt \\ &= \int_S \left(\int_{\Psi_t(W)} f(t, \mathbf{z}) d\mathbf{z} \right) dt, \end{aligned} \quad (9.5)$$

where the last equality follows from the induction hypothesis. Since X is compact, there exists a compact interval $Z \subset \mathbb{R}^{n-1}$ such that $\Psi(X) \subset S \times Z$. Thus, we have

$$\begin{aligned} \int_S \left(\int_{\Psi_t(W)} f(t, \mathbf{z}) d\mathbf{z} \right) dt &= \int_S \left(\int_Z \mathbb{1}_{\Psi_t(W)}(\mathbf{z}) f(t, \mathbf{z}) d\mathbf{z} \right) dt \\ &= \int_S \left(\int_Z \mathbb{1}_{\Psi(X)}(t, \mathbf{z}) f(t, \mathbf{z}) d\mathbf{z} \right) dt \\ &= \int_Y f. \end{aligned} \quad (9.6)$$

Combining (9.5) and (9.6) gives the desired formula. $\square$

This leads to the following corollary.

Corollary 9.5.6. *Given the hypothesis of Theorem 8.7.5, if $X \subset [\mathbf{a}, \mathbf{b}]$ is a measurable subset of a compact interval and Ψ has the form given in Lemma 9.5.5, then the change of variables formula (9.3) holds for all $f \in L^1(Y; \mathbb{R})$.*

Proof. See Exercise 9.22. $\square$

Lemma 9.5.7. *Assume that $U, V, W \subset \mathbb{R}^n$ are open and $\Psi : U \rightarrow V$ and $\Phi : V \rightarrow W$ are diffeomorphisms. Let $X \subset U$ be measurable, with $Y = \Psi(X)$ and $Z = \Phi(Y)$. If the change of variables formula with respect to Ψ holds for all $g \in L^1(Y; \mathbb{R})$ and the change of variables formula with respect to Φ holds for some $f \in L^1(Z; \mathbb{R})$, then the formula holds with respect to $\Phi \circ \Psi$ for f .*

Proof. The proof is Exercise 9.23. $\square$

Lemma 9.5.8. *If $\Psi : U \rightarrow V$ is a diffeomorphism and $X \subset U$ is a measurable subset with $Y = \Psi(X)$, then for any $\mathbf{x} \in X$ there exists an open neighborhood $W \subset U$ of $\mathbf{x}$ such that the statement of the change of variables theorem holds for every integrable function on $\Psi(X \cap W) = Y \cap \Psi(W)$.*

Proof. First we show that there is an open neighborhood W of $\mathbf{x}$ where the diffeomorphism Ψ is a composition of diffeomorphisms of the form $\Psi(\mathbf{t}) = \Psi(t_1, \dots, t_n) = (t_i, \Psi_2(\mathbf{t}), \dots, \Psi_n(\mathbf{t}))$ for some i , or of the form $\Psi(\mathbf{t}) = (\Psi_1, \dots, \Psi_{n-1}(\mathbf{t}), t_i)$. For each i let $C_{1,i}$ be the $(1, i)$ cofactor of $D\Psi(\mathbf{x})$, as given in Definition 2.9.11. Using the cofactor expansion in the first row (Theorem 2.9.16), we have

$$0 \neq \det(D\Psi) = \sum_{j=1}^n \frac{\partial \Psi_1}{\partial x_j}(\mathbf{x}) C_{1,j},$$

so there must be some i such that $C_{1,i} \neq 0$. Choose one such i and define $\psi : U \rightarrow \mathbb{R}^n$ by $\psi(\mathbf{t}) = (t_i, \Psi_2, \dots, \Psi_n)$. Note that $|\det(D\psi(\mathbf{x}))| = |C_{1,i}| \neq 0$, so by the inverse function theorem (Theorem 7.4.8), there exists an open neighborhood $W' \subset U$ of $\mathbf{x}$ such that ψ has a C^1 inverse on $\psi(W') \subset V$. In particular, ψ is a diffeomorphism on W' of the required form.

Let $\Phi : \psi(W') \rightarrow V$ be the diffeomorphism $\Phi = \Psi \circ \psi^{-1}$, so that $\Psi = \Phi \circ \psi$. Letting $\mathbf{z} = (z_1, \dots, z_n) = \psi(t_1, \dots, t_n) = (t_i, \Psi_2(\mathbf{t}), \dots, \Psi_n(\mathbf{t}))$, we have

$$\Phi(z_1, \dots, z_n) = (\Psi(\psi^{-1}(\mathbf{z})), z_2, \dots, z_n),$$

so Φ is also of the required form.

Since W' is an open neighborhood of $\mathbf{x}$, there is a compact interval $[a, b] \subset W'$ with nonempty interior $W \subset [a, b]$ such that $\mathbf{x} \in W$. By Corollary 9.5.6 and Lemma 9.5.7, the change of variables formula holds on $X \cap W$. $\square$

Now we can prove the full version of the change of variables formula.

Proof of Theorem 8.7.5. It suffices to prove the theorem for the case of $X = U$ and $f \in L^1(V; \mathbb{R})$. For each $\ell \in \mathbb{Z}^+$ let

$$K_\ell = \{\mathbf{t} \in U \mid \text{dist}(\mathbf{t}, U^c) \geq 1/\ell \text{ and } |\mathbf{t}| \leq \ell\}.$$

Each K_ℓ is a compact subset of U , and $K_\ell \subset K_{\ell+1}$. Moreover, we have $\bigcup_{\ell=1}^{\infty} K_\ell = U$. For each ℓ and for each $\mathbf{t} \in K_\ell$, Lemma 9.5.8 gives an open neighborhood $W_{\mathbf{t}}$ of $\mathbf{t}$ such that the change of variables theorem holds for all functions in $L^1(\Psi(W_{\mathbf{t}}); \mathbb{R})$.

Since K_ℓ is compact, there is a finite subcollection of the open sets $W_{\mathbf{t}}$ that covers K_ℓ . Thus, there is a countable collection $W_1, W_2, \dots$ of open subsets of U such that the change of variables theorem holds for f on each $\Psi(W_i)$, and $U = \bigcup_{i=1}^{\infty} W_i$. For each i let $W'_i = W_i \setminus \bigcup_{j=1}^{i-1} W_j$. All the W'_i are disjoint and $U = \bigcup_{i=1}^{\infty} W'_i$. Also, each W'_i is a measurable subset of W_i .

If f is nonnegative a.e., then because countable sums commute with integration (Exercise 8.16) we have

$$\begin{aligned} \int_U f \circ \Psi |\det(D\Psi)| &= \sum_{i=1}^{\infty} \int_U \mathbb{1}_{W'_i} f \circ \Psi |\det(D\Psi)| = \sum_{i=1}^{\infty} \int_{W_i} \mathbb{1}_{W'_i} f \circ \Psi |\det(D\Psi)| \\ &= \sum_{i=1}^{\infty} \int_{W_i} (\mathbb{1}_{\Psi(W'_i)} f) \circ \Psi |\det(D\Psi)| = \sum_{i=1}^{\infty} \int_{W_i} \mathbb{1}_{\Psi(W'_i)} f \\ &= \sum_{i=1}^{\infty} \int_{\Psi(W'_i)} f = \int_{\Psi(U)} f. \end{aligned}$$

So the formula holds for nonnegative integrable functions. Writing $f = f^+ - f^-$ as a difference of nonnegative integrable functions shows that the formula holds for any f . $\square$

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

9.1. Prove Proposition 9.1.1.

9.2. Prove the claim in the proof of Theorem 9.1.2 that the function $\|\cdot\|_{X'}$ defined by $\|(\mathbf{x}_k)_{k=0}^{\infty}\|_{X'} = \lim_{k \rightarrow \infty} \|\mathbf{x}_k\|$ is a seminorm. Provide an example of a sequence $(\mathbf{x}_k)_{k=0}^{\infty}$ such that $\|(\mathbf{x}_k)_{k=0}^{\infty}\| = 0$ but $(\mathbf{x}_k)_{k=0}^{\infty} \neq (0)_{k=1}^{\infty}$ (the zero element of X').

9.3. Prove Proposition 9.1.7.

9.4. Show that every element $f \in L^1([a, b]; \mathbb{C})$ can be written as $f = u + iv$ where $u, v \in L^1([a, b]; \mathbb{R})$, and conversely, for any $u, v \in L^1([a, b]; \mathbb{R})$, we have $u + iv \in L^1([a, b]; \mathbb{C})$.

- 9.5. For each $\ell \in \mathbb{N}$ let $Q_\ell \subset \mathbb{R}^n$ be the lattice of points with coordinates all lying in $2^{-\ell}\mathbb{Z}$. Show that every element of $L^1([a, b]; X)$ is equivalent to an L^1 -Cauchy sequence $(s_k)_{k=0}^\infty$ of step functions where all the corners of all the intervals involved in the step function are points of Q_ℓ for some ℓ ; that is, for each $k \in \mathbb{N}$ there exists an $\ell_k \in \mathbb{N}$ such that each interval R_I appearing in the sum $s_k = \sum_{I \in \mathcal{P}} \mathbf{x}_I \mathbb{1}_{R_I}$ is of the form $R_I = [a_I, b_I]$ with $a_I, b_I \in Q_{\ell_k}$.
-
- 9.6. Define a *generalized step function* to be a function f of the form $f = \sum_{i=1}^m c_i \mathbb{1}_{R_i}$ such that each R_i is any interval (possibly all open, all closed, all partially open, or any combination of these).
 Prove that in Definition 9.2.1 we may use generalized step functions with all open intervals. That is, prove that a set E has measure zero in the sense of Definition 9.2.1 if and only if there is an L^1 -Cauchy sequence $(f_k)_{k=0}^\infty$ of generalized step functions of the form $f_k = \sum_{i=1}^{m_k} c_{k,i} \mathbb{1}_{R_{k,i}}$ such that each $R_{k,i}$ is an open interval and $|f_k(t)| \rightarrow \infty$ for every $t \in E$. What changes if we use generalized step functions with all closed intervals?
- 9.7. Prove Lemma 9.2.2 for a monotone decreasing sequence of functions.
- 9.8. Prove Proposition 9.2.6.
- 9.9. Prove that if $I \subset \mathbb{R}^n$ is an interval and $f \in \mathcal{R}(I; \mathbb{R})$, then the graph $\Gamma_f = \{(\mathbf{x}, f(\mathbf{x})) \subset I \times \mathbb{R}\} \subset \mathbb{R}^{n+1}$ has measure zero.
-
- 9.10. Find a sequence $(f_k)_{k=0}^\infty$ of functions in $\mathcal{R}([0, 1]; \mathbb{R})$ such that $f_k \rightarrow 0$ a.e. but $\lim_{k \rightarrow \infty} \int_0^1 f_k(t) dt \neq 0$.
- 9.11. Let f be continuous on $[a, b] \subset \mathbb{R}$. Describe in detail how to construct a monotone increasing sequence $(s_k)_{k=0}^\infty$ of step functions such that $s_k \rightarrow f$ a.e. and $\lim_{k \rightarrow \infty} \int_{[a, b]} s_k < \infty$.
- 9.12. Prove Proposition 8.4.2(iii).
- 9.13. Give an example of a function on $\mathbb{R}$ that is measurable but is not integrable.
- 9.14. Let f and g be measurable. Prove the following:
- (i) The set of measurable functions is a vector space.
 - (ii) The product fg is measurable, and a product of any finite number of measurable functions is measurable.
 - (iii) If $(f_k)_{k=0}^\infty$ is a sequence of measurable functions that converges to f almost everywhere, then f is measurable.
- 9.15. Prove Proposition 9.3.7.
- 9.16. Let $f \in L^1([a, b]; \mathbb{R})$. Show that for every function $g \in \mathcal{R}([a, b]; \mathbb{R})$ the product fg lies in $L^1([a, b]; \mathbb{R})$.
- 9.17. Prove that if f is any measurable function and $g \in L^1([a, b]; \mathbb{R})$ is such that $|f| \leq g$ a.e., then $f \in L^1([a, b]; \mathbb{R})$. Hint: Generalize the arguments in the proof of Proposition 8.4.2.
-
- 9.18. Prove Fubini's theorem for step functions (Proposition 9.4.1).

- 9.19. Suppose that $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ and both $f: X \rightarrow \mathbb{R}$ and $g: Y \rightarrow \mathbb{R}$ are integrable. Prove that the function $h: X \times Y \rightarrow \mathbb{R}$, given by $h(x, y) = f(x)g(y)$, is integrable and

$$\int_{X \times Y} h = \int_X f \int_Y g.$$

- 9.20. Prove Proposition 8.6.5.

9.21. Prove Corollary 9.5.3.

9.22. Prove Corollary 9.5.6.

9.23. Prove Lemma 9.5.7. Hint: Use the chain rule and the fact that determinants are multiplicative ($\det(AB) = \det(A)\det(B)$).

Notes

As mentioned in the notes at the end of the previous chapter, much of our development of integration is inspired by [Cha95]. Other references are given at the end of the previous chapter. The proof of Proposition 9.2.5 is modeled after [Cha95, Sect. II.2.4], and the proof of Proposition 9.3.3 is from [Cha95, Sect. II.2.3]. The proof of Proposition 9.3.4 is from [Cha95, Sect. II.3.3]. The proof of the change of variables formula is based on [Mun91, Sect. 19] and [Dri04, Sect. 20.2].

10 Calculus on Manifolds

Cultivate your curves—they may be dangerous, but they won't be avoided.
—Mae West

In this chapter we describe how to compute tangents, and integrals on *parametrized manifolds* in a Banach space. Manifolds are curves, surfaces, and other spaces that are, at least locally, images of especially nice, smooth maps from an open set in $\mathbb{R}^k$ into some Banach space. We begin with manifolds whose domain is one dimensional, that is, smooth curves. We describe how to compute tangents, normals, arclength, and more general line integrals that account for the geometry of the curve. These are important tools for complex analysis and spectral calculus. We then give a brief treatment of the theory for more general manifolds.

The main result of this chapter is Green's theorem, which is a generalization of the fundamental theorem of calculus. It tells us that under certain conditions the line integral around a simple closed curve in the plane is the same as the integral of a certain derivative of the integrand on the region bounded by the curve. In other words, the surface integral of the derivative on the region is equal to the line integral of the original integrand on the boundary of the region.

10.1 Curves and Arclength

In the next two sections we discuss smooth curves. Although these may seem like very simple manifolds, much of the general theory is motivated by this basic case. Throughout this section, assume that $(X, \|\cdot\|)$ is a Banach space.

Definition 10.1.1. A smooth parametrized curve in X is an injective C^1 function $\sigma : I \rightarrow X$, where $I \subset \mathbb{R}$ is an interval (not necessarily closed, open, or finite) such that σ' never vanishes. It is common to call the domain of a parametrization “time” and to use the variable t to denote points in I . If the domain I of σ is not an open interval, then C^1 is meant in the sense of Remark 7.2.5. We say that $\sigma : [a, b] \rightarrow X$ is a simple closed parametrized curve if $\sigma(a) = \sigma(b)$ and σ is injective when restricted to $[a, b)$.

Remark 10.1.2. With some work it is possible to show the results of this chapter also hold for curves $\sigma : I \rightarrow X$ on a closed interval I that are not differentiable at the endpoints of I , but that are continuous on I and C^1 on the interior of I .

Unexample 10.1.3.

- (i) The derivative of the map $\alpha : \mathbb{R} \rightarrow \mathbb{R}^2$ given by $\alpha(t) = (1 - t^2, t^3 - t)$ never vanishes, but α does not define a smooth curve because it is not injective—it takes on the value $(0, 0)$ at $t = 1$ and $t = -1$.
- (ii) The map $\alpha : \mathbb{R} \rightarrow \mathbb{R}^2$ given by $\alpha(t) = (t^2, t^3)$ is injective, but it is not a smooth curve because the derivative $D\alpha$ vanishes at $t = 0$.

We define the *tangent* to the curve σ at time t_1 to be the vector $\sigma'(t_1)$. If a curve in $\mathbb{R}^n$ is thought of as the trajectory of a particle, then $\sigma'(t_1)$ is its velocity at time t_1 . The line in X that is tangent to the curve at time t_1 is defined by the parametrization $L(t) = t\sigma'(t_1) + \sigma(t_1)$.

10.1.1 Parametrizations and Equivalent Curves

We often want to study the underlying curve itself, that is, the image of the map σ , rather than the parametrization of the curve. In other words, we want the objects we study to be independent of parametrization.

Definition 10.1.4. Two smooth parametrized curves $\sigma_1 : I \rightarrow X$ and $\sigma_2 : J \rightarrow X$ are equivalent if there exists a bijective C^1 map $\phi : I \rightarrow J$, such that $\phi'(t) > 0$ for all $t \in I$ and $\sigma_2 \circ \phi = \sigma_1$. In this case, we say that σ_2 is a reparametrization of σ_1 . Each equivalence class of parametrizations is called a smooth, oriented curve.

If we replace the condition that $\phi'(t) > 0$ by the condition that $\phi'(t) \neq 0$ for all t , we get a larger equivalence class that includes orientation-reversing reparametrizations. Each of these larger equivalence classes is called a smooth, unoriented curve or just a smooth curve.

Remark 10.1.5. The tangent vector $\sigma'(t_1)$ to the curve σ at the point $\sigma(t_1)$ is not independent of parametrization. A reparametrization $\phi : I \rightarrow J$ scales the tangent vector by ϕ' . However, we can define the *unit tangent* $\mathbf{T}$ to be

$$\mathbf{T}(t) = \sigma'(t) / \|\sigma'(t)\|,$$

and the unit tangent at each point depends only on the orientation of the curve. See Figure 10.1.

Definition 10.1.6. A finite collection of smooth parametrized curves $\sigma_1 : [a_1, b_1] \rightarrow X, \dots, \sigma_k : [a_k, b_k] \rightarrow X$ is called a piecewise-smooth parametrized curve if we have $\sigma_i(b_i) = \sigma_{i+1}(a_{i+1})$ for each $i \in \{1, \dots, k-1\}$. Such a curve is often denoted $\sigma_1 + \dots + \sigma_k$.

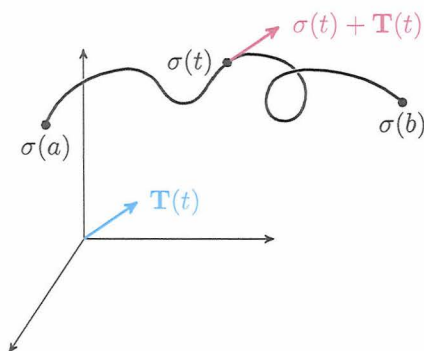


Figure 10.1. A smooth parametrized curve σ with starting point $\sigma(a)$ and ending point $\sigma(b)$, as given in Definition 10.1.1. For each $t \in (a, b)$, the vector $\mathbf{T}(t)$ is the unit tangent to the curve at time t . See Remark 10.1.5 for the definition of $\mathbf{T}(t)$.

Remark 10.1.7. Throughout this section we focus primarily on smooth parametrized curves, but it should be clear how to extend these results to piecewise-smooth curves.

10.1.2 Arclength

Arclength measures the length of an oriented curve in a way that is independent of parametrization. It is essential for understanding the geometry of a curve. It should seem intuitive that arclength would be the integral of speed, where speed is the norm $\|\sigma'(u)\|$ of the velocity.

Definition 10.1.8. The arclength $\text{len}(\sigma)$ of a smooth parametrized curve $\sigma : [a, b] \rightarrow X$ is

$$\text{len}(\sigma) = \int_a^b \|\sigma'(u)\| du.$$

Example 10.1.9. Consider a straight line segment, parametrized by $\sigma : [a, b] \rightarrow X$ with $\sigma(u) = u\mathbf{x} + \mathbf{v}$, where $\mathbf{x}, \mathbf{v} \in X$ are fixed. The traditional definition of the length of the segment is just the norm of the difference between the starting and ending points: $\|a\mathbf{x} + \mathbf{v} - (b\mathbf{x} + \mathbf{v})\| = (b - a)\|\mathbf{x}\|$. It is immediate to check that this agrees with our new definition of arclength.

Example 10.1.10. Consider a circle in $\mathbb{R}^2$ of radius r , centered at the origin, with parametrization $\sigma : [0, 2\pi] \rightarrow \mathbb{R}^2$ given by $\sigma(u) = (r \cos(u), r \sin(u))$. The length of the curve is

$$\int_0^{2\pi} \|\sigma'(u)\| du = \int_0^{2\pi} r \|(-\sin(u), \cos(u))\| du = 2\pi r,$$

which agrees with the classical definition of arclength of the circle.

Example 10.1.11. The graph of a function $f \in C^1([a, b]; \mathbb{R})$ defines a smooth parametrized curve $\sigma : [a, b] \rightarrow \mathbb{R}^2$ by $\sigma(t) = (t, f(t))$. We have $\sigma'(t) = (1, f'(t))$ so that

$$\text{len}(\sigma) = \int_a^b \sqrt{1 + (f'(u))^2} du.$$

Proposition 10.1.12. *Arclength is independent of parametrization. In other words, if $\sigma : [a, b] \rightarrow X$ and $\gamma = \sigma \circ \phi$ where $\phi : [c, d] \rightarrow [a, b]$ is a bijective C^1 function with $\phi'(t) \neq 0$ for all $t \in [c, d]$, then $\text{len}(\sigma) = \text{len}(\gamma)$.*

Proof. Since $\phi'(t) \neq 0$ for all $t \in [c, d]$ the intermediate value theorem guarantees that either $\phi'(t) > 0$ for all t or $\phi'(t) < 0$ for all t . We prove the case of $\phi'(t) < 0$. The case of $\phi'(t) > 0$ is similar (and slightly easier).

Since ϕ is bijective, it must map some point t_0 to b . If $t_0 > c$, then the mean value theorem guarantees that for some $\xi \in [c, t_0]$ we have

$$b - \phi(c) = \phi'(\xi)(t_0 - c) < 0,$$

which is impossible, since $\phi(c) \in [a, b]$; thus, $\phi(c) = b$. A similar argument shows that $\phi(d) = a$. The substitution $u = \phi(t)$ now gives

$$\begin{aligned} \text{len}(\gamma) &= \int_c^d \|\gamma'(t)\| dt = \int_c^d \|\sigma'(\phi(t))\| |\phi'(t)| dt \\ &= - \int_b^a \|\sigma'(u)\| du = \text{len}(\sigma). \quad \square \end{aligned}$$

Definition 10.1.13. *Given $\sigma : [a, b] \rightarrow X$, define the arclength function $s : [a, b] \rightarrow \mathbb{R}$ by assigning $t \in [a, b]$ to the length of σ restricted to the subinterval $[a, t]$*

$$s(t) = \text{len}(\sigma|_{[a, t]}).$$

If $a = 0$ and $\|\sigma'(u)\| = 1$ for all u in $[0, b]$, then $s(t) = t$, and we say that σ is parametrized by arclength.

Example 10.1.14. If $\sigma : [0, \infty) \rightarrow \mathbb{R}^3$ is the helix given by

$$\sigma(t) = (t, \cos(t), \sin(t)),$$

then

$$s(t) = \int_0^t \sqrt{1 + \sin^2(t) + \cos^2(t)} dt = t\sqrt{2},$$

so σ is not parametrized by arclength, but it is easy to fix this. Setting $t = \tau/\sqrt{2}$ gives a new parametrization

$$\alpha(\tau) = \sigma(t(\tau)) = (\tau/\sqrt{2}, \cos(\tau/\sqrt{2}), \sin(\tau/\sqrt{2})),$$

with $s(t(\tau)) = \tau$ and

$$\begin{aligned} \|\alpha'(\tau)\| &= \left\| \frac{d\sigma(t(\tau))}{d\tau} \right\| = \frac{1}{\sqrt{2}} \|(1, -(\sin(\tau/\sqrt{2})), (\cos(\tau/\sqrt{2})))\| \\ &= \frac{1}{\sqrt{2}} \sqrt{1 + (\sin^2(\tau/\sqrt{2}) + \cos^2(\tau/\sqrt{2}))} = 1. \end{aligned}$$

So $\alpha(\tau)$ is parametrized by arclength.

The previous example is not a fluke—we can always reparametrize a smooth curve so that it is parametrized by arclength, as the next two propositions show.

Proposition 10.1.15. *For any smooth parametrized curve $\sigma : [a, b] \rightarrow X$ the function $s(t)$ has an inverse $\rho : [0, L] \rightarrow [a, b]$, where $L = \text{len}(\sigma) = s(b)$.*

Proof. The proof is Exercise 10.2. $\square$

Definition 10.1.16. *Given any smooth curve with parametrization σ we define a new parametrization $\gamma : [0, L] \rightarrow X$ by $\gamma(s) = \sigma \circ \rho(s)$. We call γ the parametrization of the curve by arclength.*

Proposition 10.1.17. *For any smooth curve with parametrization σ the parametrization $\gamma = \sigma \circ \rho$ of the curve is, in fact, a parametrization by arclength; that is,*

$$\|\gamma'(s)\| = 1 \quad \text{for all } s \in [0, L].$$

Proof. This follows immediately from the fact that $s'(t) = \|\sigma'(t)\|$ combined with the chain rule. $\square$

Example 10.1.18. The circle $\sigma : [0, 2\pi] \rightarrow \mathbb{R}^2$ with $\sigma(t) = (r \cos(t), r \sin(t))$ in Example 10.1.10 is parametrized by arclength only if $|r| = 1$. To see this, note that the arclength function is $s(t) = \int_0^t \|\sigma'(u)\| du = tr$, which has inverse $\rho(s) = s/r$. Thus, $\gamma(s) = (r \cos(s/r), r \sin(s/r))$ is the parametrization of this circle by arclength.

10.2 Line Integrals

In this section we define the line integral of a function f over a curve. We want the line integral to be independent of the parametrization and, in analogy to the traditional single-variable integral, to correspond to summing the value of the function on a short segment of the curve times the length of the segment. A physical interpretation of this would correspond to the mass of the curve if the function f describes the density of the curve at each point.

Throughout this section, assume that $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces over the same field $\mathbb{F}$.

10.2.1 Line Integrals

Definition 10.2.1. Given a smooth curve $C \subset X$ parametrized by arclength $\gamma : [0, L] \rightarrow C$ and a function $f : C \rightarrow Y$, we define the line integral of f over C to be

$$\int_C f ds = \int_0^L f(\gamma(s)) ds,$$

if the integral exists.

Proposition 10.2.2. Given a smooth curve C with parametrization $\sigma : [a, b] \rightarrow C$ (not necessarily parametrized by arclength) and a function $f : C \rightarrow Y$, the line integral of f over C can be computed as

$$\int_C f ds = \int_a^b f(\sigma(t)) \|\sigma'(t)\|_X dt,$$

if that integral exists.

Proof. The proof is Exercise 10.8. $\square$

Remark 10.2.3. Since every oriented curve has only one parametrization by arclength, the line integral depends only on the oriented curve class of C .

Example 10.2.4. Suppose $C \subset \mathbb{R}^3$ is given by the parametrization

$$\sigma(t) = (2 \sin(t), t, -2 \cos(t)),$$

where $0 \leq t \leq \pi$. The line integral $\int_C xyz ds$ can be evaluated as

$$\begin{aligned} - \int_0^\pi 4t \sin(t) \cos(t) \|\sigma'(t)\| dt &= -\sqrt{5} \int_0^\pi 4t \sin(t) \cos(t) dt \\ &= \sqrt{5} (t \cos(2t) - \sin(t) \cos(t)) \Big|_0^\pi = \sqrt{5}\pi. \end{aligned}$$

Example 10.2.5.

- (i) If $C \subset X$ is parametrized by $\sigma : [a, b] \rightarrow C$, then the line integral $\int_C ds = \int_a^b \|\sigma'(t)\|_X dt$ is just the arclength of C .
- (ii) If $\rho : C \rightarrow \mathbb{R}$ is positive for all $c \in C$, we can interpret it as giving the density of a wire in the shape of the curve C . In this case, as mentioned in the introduction to this section, the integral $\int_C \rho ds$ is the mass of C .
- (iii) The center of mass of a wire in the shape of the curve C with density function ρ is $(\bar{x}_1, \dots, \bar{x}_n)$, where

$$\bar{x}_k = \frac{1}{m} \int_C x_k \rho ds,$$

and $m = \int_C \rho ds$ is the mass.

10.2.2 Line Integrals of Vector Fields

In the special case where C is a curve in $X = \mathbb{F}^n$, we call a function $F : C \rightarrow \mathbb{F}^n$ a *vector field on C* . For vector fields it is often productive to consider a line integral that takes into account not only the length of each infinitesimal curve segment, but also its direction. This integral is useful, for example, when computing the work done by a force F moving a particle along C . If $F \in \mathbb{R}^n$ is a constant force, the work done moving a particle along a line segment of the form $t\mathbf{v} + \mathbf{c}$, where $\mathbf{v} \in \mathbb{R}^n$ and $t \in [a, b]$, is given by

$$\text{work} = F \cdot (b - a)\mathbf{v} = (b - a) \langle F, \mathbf{v} \rangle.$$

If C is parametrized by $\sigma : [a, b] \rightarrow \mathbb{R}^n$ and the force $F : C \rightarrow \mathbb{R}^n$ is not necessarily constant, then over a small interval $\Delta\sigma$ of the curve containing $\sigma(t)$, the work done is approximately $F(\sigma(t)) \cdot \Delta\sigma$. Summing these pieces and taking the limit as $\Delta\sigma \rightarrow 0$ yields an integral giving the work done by the force F to move a particle along the curve C . This motivates the following definition.

Definition 10.2.6. Given a curve $C \subset \mathbb{F}^n$ with parametrization $\sigma : [a, b] \rightarrow C$ and a C^1 vector field $F : C \rightarrow \mathbb{F}^n$, if $F(\sigma(t)) \cdot \sigma'(t)$ is integrable on $[a, b]$, then we define the line integral of the vector field F over C to be

$$\int_C F \cdot d\sigma = \int_C F \cdot \mathbf{T} ds = \int_a^b F(\sigma(t)) \cdot \sigma'(t) dt, \quad (10.1)$$

where $\mathbf{T}$ is the unit tangent to C . The $\cdot$ in the left-hand integral is a formal symbol defined by (10.1), while the $\cdot$ in the second and third integrals is the usual dot-product: $\mathbf{x} \cdot \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y}$.

If we write σ and F in terms of the standard basis as

$$\sigma(t) = \mathbf{x}(t) = (x_1(t), \dots, x_n(t))$$

and $F(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_n(\mathbf{x}))$, then it is traditional to write $dx_i = x'_i(t) dt$, and in this notation we have

$$\int_C F \cdot d\sigma = \int_a^b \sum_{i=1}^n F_i(\mathbf{x}(t)) x'_i(t) dt = \sum_{i=1}^n \int_C F_i(\mathbf{x}) dx_i.$$

Remark 10.2.7. Using this new notation, our previous discussion tells us that the line integral

$$\int_C F \cdot d\sigma$$

is the work done by a force $F : C \rightarrow \mathbb{R}^n$ moving a particle along the curve C .

Proposition 10.2.8. *The line integral of a vector field F over a smooth curve does not depend on the parametrization of the curve. That is, given two equivalent parametrizations $\sigma_1 : [a, b] \rightarrow C$ and $\sigma_2 : [c, d] \rightarrow C$, we have*

$$\int_a^b F(\sigma_1(t)) \cdot \sigma'_1(t) dt = \int_c^d F(\sigma_2(t)) \cdot \sigma'_2(t) dt.$$

However, if a reparametrization changes the orientation, then it changes the sign of the integral.

Proof. The proof is Exercise 10.9. $\square$

Example 10.2.9. If C is a segment of a helix parametrized by $\sigma : [0, 2\pi] \rightarrow \mathbb{R}^3$, with $\sigma(t) = (\cos(t), \sin(t), t)$, and if $F(x, y, z) = (-y, x, z^2)$, then

$$\begin{aligned} \int_C F \cdot d\sigma &= \int_C -y dx + \int_C x dy + \int_C z^2 dz \\ &= \int_0^{2\pi} \sin^2(t) dt + \int_0^{2\pi} \cos^2(t) dt + \int_0^{2\pi} t^2 dt \\ &= 2\pi + \frac{8\pi^3}{3}. \end{aligned}$$

Example 10.2.10. If $F = D\phi$ is the derivative of a scalar-valued function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ (in this context ϕ is usually called a *potential*), then we say that F is a *conservative vector field*. In this case the fundamental theorem of calculus gives

$$\begin{aligned} \int_C F \cdot d\sigma &= \int_a^b F(\sigma(t)) \cdot \sigma'(t) dt = \int_a^b D\phi \cdot \sigma'(t) dt \\ &= \int_a^b D(\phi \circ \sigma) dt = \phi(\sigma(b)) - \phi(\sigma(a)). \end{aligned}$$

In this case the integral depends only on the value of the potential at the endpoints—it is independent of the path C and of σ . This is an important phenomenon that we revisit several times.

10.3 Parametrized Manifolds

A parametrized manifold is the natural higher-dimensional generalization of the concept of a parametrized curve. Throughout this section assume that $(X, \|\cdot\|)$ is a Banach space.

Definition 10.3.1. Let U be an open subset of $\mathbb{R}^m$. We say that $\alpha \in C^1(U; X)$ is a parametrized m -manifold if it is injective and at each point $\mathbf{u} \in U$ the derivative $D\alpha(\mathbf{u})$ is injective (that is, $D\alpha(\mathbf{u})$ has rank m). A parametrized 2-manifold is also called a parametrized surface.

Remark 10.3.2. The injectivity of $D\alpha$ implies that $m \leq \dim(X)$.

Example 10.3.3.

- (i) Every smooth parametrized curve with an open domain is a parametrized 1-manifold.
- (ii) For every $n \in \mathbb{Z}^+$ the identity map $I : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a parametrized n -manifold.
- (iii) For any $U \subset \mathbb{R}^m$, the graph of any C^1 function $f : U \rightarrow \mathbb{R}$ gives a parametrized m -manifold in $\mathbb{R}^{m+1}$ by $\alpha(\mathbf{u}) = (\mathbf{u}, f(\mathbf{u})) \in \mathbb{F}^{m+1}$. (Check that $D\alpha$ has rank m .)

10.3.1 Parametrizations and Equivalent Manifolds

As with curves, we often want to study the underlying spaces (manifolds) themselves, rather than the parametrizations; that is, we want to understand properties that are independent of parametrization.

Definition 10.3.4. Two parametrized m -manifolds $\alpha_1 : U_1 \rightarrow X$ and $\alpha_2 : U_2 \rightarrow X$ are equivalent if there exists a bijective C^1 map $\phi : U_1 \rightarrow U_2$, such that

- (i) ϕ^{-1} is also C^1 ,
- (ii) $\det(D\phi(\mathbf{u})) > 0$ for all $\mathbf{u} \in U_1$, and
- (iii) $\alpha_2 \circ \phi = \alpha_1$.

In this case, we say that α_2 is a reparametrization of α_1 . Each equivalence class of parametrizations is called an oriented m -manifold or, if m is understood, it is just called an oriented manifold.

If we drop the condition $\det(D\phi) > 0$, then we get a larger equivalence class of manifolds. Since the derivative $D\phi$ is continuous and nonvanishing, it is either always positive or always negative. If $\det(D\phi(\mathbf{u})) < 0$ for all $\mathbf{u} \in U_1$, then we say that ϕ is orientation reversing. Each of these larger equivalence classes is called an unoriented m -manifold or just an m -manifold.

Example 10.3.5. We may parametrize the upper half of the unit sphere $S^2 \subset \mathbb{R}^3$ by $\alpha : U \rightarrow \mathbb{R}^3$, where $U = B(0, 1)$ is the unit disk in the plane and $\alpha(u, v) = (u, v, \sqrt{1 - u^2 - v^2})$. But we may also use the parametrization $\sigma(u, v) = (v, u, \sqrt{1 - u^2 - v^2})$. These are equivalent by the map $G : U \rightarrow U$ given by $G(u, v) = (v, u)$. Moreover, we have

$$\det(DG) = \det \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -1,$$

so the equivalence reverses the orientation.

We may also parametrize the upper half of the unit sphere by $\beta : (0, \pi) \times (0, \pi) \rightarrow \mathbb{R}^3$ where $\beta(\phi, \theta) = (\cos(\phi), \cos(\theta) \sin(\phi), \sin(\theta) \sin(\phi))$. The parametrizations α and β are equivalent by the map $F : (0, \pi) \times (0, \pi) \rightarrow B(0, 1)$, given by $F(\phi, \theta) = (\cos(\phi), \cos(\theta) \sin(\phi))$. This map is C^∞ and bijective with inverse $F^{-1}(u, v) = (\text{Arccos}(u), \text{Arccos}(v/\sqrt{1 - u^2}))$, which is also C^∞ . Moreover, we have

$$\det(DF) = \det \begin{bmatrix} -\sin(\phi) & 0 \\ \cos(\phi) \cos(\theta) & -\sin(\phi) \sin(\theta) \end{bmatrix} = \sin^2(\phi) \sin(\theta) > 0,$$

so this equivalence preserves orientation.

10.3.2 Tangent Spaces and Normals

As in the case of curves, the derivative $D\alpha$ is not independent of parametrization, so if α and β are two parametrizations of M , the derivatives $D\alpha$ and $D\beta$ are usually not equal. However, the image of the derivative—the *tangent space*—is the same for all equivalent parametrizations.

Definition 10.3.6. Given a parametrized m -manifold $\alpha : U \rightarrow M \subset X$, and a point $\mathbf{u} \in U$ with $\alpha(\mathbf{u}) = \mathbf{p}$, define the tangent space $T_{\mathbf{p}}M$ of M at $\mathbf{p}$ to be the image of the derivative $D\alpha(\mathbf{u}) : \mathbb{R}^m \rightarrow X$ in X :

$$T_{\mathbf{p}}M = \mathcal{R}(D\alpha(\mathbf{u})).$$

Thus, if $\mathbf{v}_1, \dots, \mathbf{v}_m$ is any basis for $\mathbb{R}^m$, then the vectors $D\alpha(\mathbf{u})\mathbf{v}_1, \dots, D\alpha(\mathbf{u})\mathbf{v}_m$ form a basis for $T_{\mathbf{p}}M$.

Example 10.3.7. For each of the parametrizations α , β , and σ of Example 10.3.5, we compute the tangent space of $S^2 \subset \mathbb{R}^3$ at the point

$$\mathbf{p} = \left(\frac{1}{2}, 0, \frac{\sqrt{3}}{2} \right) = \alpha \left(\frac{1}{2}, 0 \right) = \sigma \left(0, \frac{1}{2} \right) = \beta \left(\frac{\pi}{3}, \frac{\pi}{2} \right).$$

We have

$$D\alpha(u, v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-u}{\sqrt{1-u^2-v^2}} & \frac{-v}{\sqrt{1-u^2-v^2}} \end{bmatrix},$$

so the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ are mapped by $D\alpha(1/2, 0)$ to $[1 \ 0 \ -1/\sqrt{3}]^T$ and $[0 \ 1 \ 0]^T$, respectively. A similar calculation shows that for the parametrization σ the standard basis vectors are mapped by $D\sigma(1/2, 0)$ to $[0 \ 1 \ 0]^T$ and $[1 \ 0 \ -1/\sqrt{3}]^T$, respectively. And for the parametrization β we have

$$D\beta = \begin{bmatrix} -\sin(\phi) & 0 \\ \cos(\theta)\cos(\phi) & -\sin(\theta)\sin(\phi) \\ \sin(\theta)\cos(\phi) & \cos(\theta)\sin(\phi) \end{bmatrix},$$

and thus the standard basis vectors are mapped by $D\beta(\pi/3, \pi/2)$ to $[-\sqrt{3}/2 \ 0 \ 1/2]^T$ and $[0 \ -\sqrt{3}/2 \ 0]^T$, respectively. It is straightforward to check that each of these pairs of vectors spans the same subspace $T_{\mathbf{p}}S^2$ of $\mathbb{R}^3$.

Proposition 10.3.8. *The tangent space $T_{\mathbf{p}}M$ is independent of parametrization and of orientation.*

Proof. Given any two parametrizations $\alpha : U \rightarrow M \subset X$ and $\beta : W \rightarrow M$ with $U, W \subset \mathbb{R}^m$ and $\alpha(\mathbf{u}) = \mathbf{p} = \beta(\mathbf{w})$, the composition $\phi = \beta^{-1} \circ \alpha$ is a reparametrization. If $\mathbf{x} \in \mathcal{R}(D\alpha(\mathbf{u}))$, then $\mathbf{x} = D\alpha(\mathbf{u})\mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^m$. At the point $\mathbf{u}$ we have

$$D(\phi)(\mathbf{u}) = D(\beta^{-1} \circ \alpha)(\mathbf{u}) = (D\beta)^{-1}(\mathbf{p})D\alpha(\mathbf{u}),$$

and so

$$\mathbf{x} = D\alpha(\mathbf{u})\mathbf{v} = D\beta(\mathbf{w})((D\beta)^{-1}(\mathbf{p})D\alpha(\mathbf{u})\mathbf{v}) \in \mathcal{R}(D\beta(\mathbf{w})).$$

This shows $\mathcal{R}(D\alpha(\mathbf{u})) \subset \mathcal{R}(D\beta(\mathbf{w}))$. The proof that $\mathcal{R}(D\beta(\mathbf{w})) \subset \mathcal{R}(D\alpha(\mathbf{u}))$ is essentially identical. $\square$

Remark 10.3.9. The tangent space $T_{\mathbf{p}}M$ to $M \subset X$ at $\mathbf{p} \in M$ is a vector subspace of X , so it always contains $\mathbf{0}$, but one often draws tangent lines or tangent planes passing through the point $\mathbf{p}$ instead through $\mathbf{0}$. To get the traditional picture of a tangent line or plane, simply add $\mathbf{p}$ to every element of $T_{\mathbf{p}}M$. See Figure 10.2 for a depiction of both $T_{\mathbf{p}}M$ and $\mathbf{p} + T_{\mathbf{p}}M$. There are many reasons to prefer $T_{\mathbf{p}}M$ over the traditional picture—perhaps the most important reason is that the traditional picture of a tangent plane is not a vector space.

In the special case of a surface in $\mathbb{R}^3$ we can use the cross product of tangent vectors to define a normal to the surface. This cross product depends strongly on the choice of parametrization, but if we rescale the normal to have length one, then the resulting *unit normal* depends only on the oriented equivalence class of the surface.

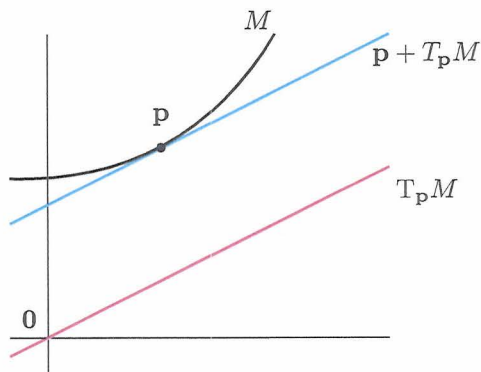


Figure 10.2. The tangent space $T_p M$ (red) to the manifold M at the point p is a vector space, and hence passes through the origin. The traditional picture that is often drawn instead is the translate $p + T_p M$ (blue). This translate is not a vector space, but there is an obvious bijection from it to the tangent space given by $v \mapsto v - p$.

Definition 10.3.10. Let $U \subset \mathbb{R}^2$ be an open set and $\alpha : U \rightarrow S \subset \mathbb{R}^3$ be a parametrized surface. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be the standard basis elements in $\mathbb{R}^2$, and for each $\mathbf{u} \in U$ we have tangent vectors $D\alpha(\mathbf{u})\mathbf{e}_1$ and $D\alpha(\mathbf{u})\mathbf{e}_2$ in $\mathbb{R}^3$.

For each $\mathbf{u} \in U$, we define the unit normal $\mathbf{N}$ to the surface S at $\alpha(\mathbf{u})$ to be

$$\mathbf{N} = \frac{D\alpha(\mathbf{u})\mathbf{e}_1 \times D\alpha(\mathbf{u})\mathbf{e}_2}{\|D\alpha(\mathbf{u})\mathbf{e}_1 \times D\alpha(\mathbf{u})\mathbf{e}_2\|},$$

where $\times$ is the cross product (see Appendix C.3). See Figure 10.3 for a depiction of the unit normal.

Proposition 10.3.11. The unit normal depends only on the orientation of the surface (that is, the orientation-preserving equivalence class). If the orientation is reversed, the unit normal changes sign: $\mathbf{N} \mapsto -\mathbf{N}$.

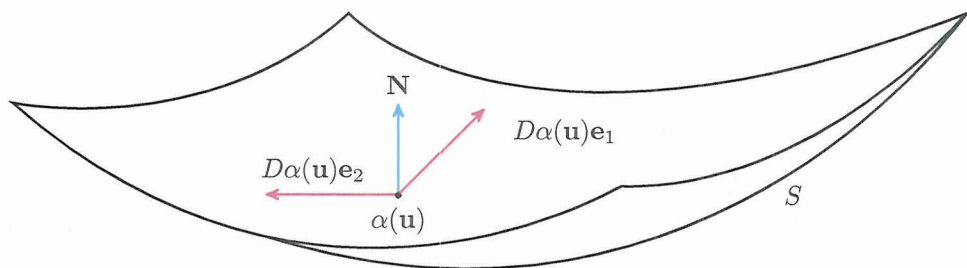


Figure 10.3. The unit normal $\mathbf{N}$ to the surface S at $\alpha(\mathbf{u})$, as described in Definition 10.3.10. The unit normal is independent of the parametrization, but if the orientation of the parametrization changes, then the sign of the unit normal is flipped; see Proposition 10.3.11.

Proof. Given any two parametrizations $\alpha : U \rightarrow M \subset \mathbb{R}^3$ and $\beta : W \rightarrow M$ with $U, W \subset \mathbb{R}^2$ and $\alpha(\mathbf{u}) = \mathbf{p} = \beta(\mathbf{w})$, such that the composition $\phi = \beta^{-1} \circ \alpha$ is orientation preserving (that is, $\det(D\phi) > 0$), and given $\mathbf{u} \in U$ we have

$$D\alpha(\mathbf{u})\mathbf{e}_i = D\beta(\phi(\mathbf{u}))D\phi(\mathbf{u})\mathbf{e}_i.$$

By Proposition C.3.2, the cross product $D\beta(\phi(\mathbf{u}))D\phi(\mathbf{u})\mathbf{e}_1 \times D\beta(\phi(\mathbf{u}))D\phi(\mathbf{u})\mathbf{e}_2$ is equal to $\det(D\phi(\mathbf{u}))(D\beta(\phi(\mathbf{u}))\mathbf{e}_1 \times D\beta(\phi(\mathbf{u}))\mathbf{e}_2)$, so we have

$$\frac{D\alpha(\mathbf{u})\mathbf{e}_1 \times D\alpha(\mathbf{u})\mathbf{e}_2}{\|D\alpha(\mathbf{u})\mathbf{e}_1 \times D\alpha(\mathbf{u})\mathbf{e}_2\|} = \frac{\det(D\phi(\mathbf{u}))}{|\det(D\phi(\mathbf{u}))|} \frac{D\beta(\phi(\mathbf{u}))\mathbf{e}_1 \times D\beta(\phi(\mathbf{u}))\mathbf{e}_2}{\|D\beta(\phi(\mathbf{u}))\mathbf{e}_1 \times D\beta(\phi(\mathbf{u}))\mathbf{e}_2\|}. \quad \square$$

Vista 10.3.12. The previous construction of the unit normal depends on the fact that we have a two-dimensional tangent space embedded in $\mathbb{R}^3$. Although the cross product is only defined for a pair of vectors in $\mathbb{R}^3$, there is a very beautiful and powerful generalization of this idea to manifolds of arbitrary dimension in X using the exterior algebra of differential forms. Unfortunately we cannot treat this topic here, but the interested reader may find a complete treatment in many books on vector calculus or differential geometry, including several listed in the notes at the end of this chapter.

10.4 * Integration on Manifolds

In the case of the line integral, we compute the length of a curve by integrating the length $\|\sigma'(t)\|$ of the tangent vector at each point of the parametrization. To compute the analogous k -dimensional volume of a k -dimensional manifold in $\mathbb{R}^n$, the correct term to integrate is the k -dimensional volume of an infinitesimal parallelepiped in the manifold.

Throughout this section we will work with manifolds in a finite-dimensional space over $\mathbb{R}$ because working over $\mathbb{C}$ or in infinite dimensions involves some additional complexities that we do not wish to address here.

10.4.1 The k -volume of a Parallelepiped in $\mathbb{R}^n$

To begin, we must understand volumes of parallelepipeds in $\mathbb{R}^n$.

Definition 10.4.1. We say that a k -dimensional parallelepiped P in $X = \mathbb{R}^n$ with one vertex at the origin is defined by vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in X$ if each of those vectors is an edge of P , and any other edge is obtained by translating one of the $\mathbf{x}_i$ to start at some sum $\sum_{\ell=1}^k \mathbf{x}_{j_\ell}$ and end at $\mathbf{x}_i + \sum_{\ell=1}^k \mathbf{x}_{j_\ell}$. Also we require that each $(k-1)$ -plane spanned by any $(k-1)$ of the vectors contains one face of P . See Figure 10.4 for an example.

Let Q be the unit interval (cube) $[0, 1] \subset \mathbb{R}^k$, where $\mathbf{1} = \sum_{i=1}^k \mathbf{e}_i$. This is the parallelepiped in $\mathbb{R}^k$ defined by the standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_k$. If $\mathbf{x}_1, \dots, \mathbf{x}_k$ is a collection of any k vectors in $\mathbb{R}^k$, we can construct a linear operator

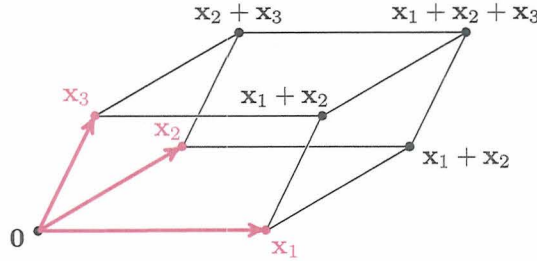


Figure 10.4. The parallelepiped defined by three vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. Each of the planes spanned by any two of these three vectors contains a face of the parallelepiped, and every edge is a translate of one of these three vectors.

$L(\mathbf{x}_1, \dots, \mathbf{x}_k) : \mathbb{R}^k \rightarrow \mathbb{R}^k$ by sending each $\mathbf{e}_i$ to $\mathbf{x}_i$. In terms of the standard basis, the matrix representation of $L(\mathbf{x}_1, \dots, \mathbf{x}_k)$ has its i th column equal to the representation of $\mathbf{x}_i$. The operator $L(\mathbf{x}_1, \dots, \mathbf{x}_k)$ maps Q to the parallelepiped defined by $\mathbf{x}_1, \dots, \mathbf{x}_k$. Therefore, the volume of this parallelepiped is equal to the modulus of the determinant of L (see Remark 8.7.6.) We can also rewrite this as

$$\text{vol}(P) = |\det(L)| = \sqrt{\det(L^T L)}.$$

A k -dimensional parallelepiped in $P \subset \mathbb{R}^n$, considered as a subset of $\mathbb{R}^n$, must have measure zero, but it is useful to make sense of the k -dimensional volume of P . We have already seen this with the length of a line segment in $\mathbb{R}^n$ —the line segment has measure zero as a subset of $\mathbb{R}^n$, but it is still useful to measure its length.

Definition 10.4.2. The k -volume of the k -dimensional parallelepiped $P \subset \mathbb{R}^n$ defined by vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ is given by

$$\text{vol}_k(P) = \sqrt{\det(L^T L)},$$

where $L : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is the linear transformation mapping each $\mathbf{e}_i \in \mathbb{R}^k$ to $\mathbf{x}_i \in \mathbb{R}^n$.

10.4.2 Integral over a k -Manifold in $\mathbb{R}^n$

If a k -manifold $M \subset \mathbb{R}^n$ is parametrized by $\alpha : U \rightarrow \mathbb{R}^n$ with $\alpha(\mathbf{u}) = \mathbf{p}$, then $D\alpha(\mathbf{u})$ maps a unit k -cube Q in U to $D\alpha(\mathbf{u})Q \subset T_{\mathbf{p}}M \subset \mathbb{R}^n$, which has k -volume equal to $\sqrt{\det(D\alpha(\mathbf{u})^T D\alpha(\mathbf{u}))}$. This motivates the following definitions.

Definition 10.4.3. Given a parametrized k -manifold $\alpha : U \rightarrow \mathbb{R}^n$, and given a measurable subset $A \subset U$ with $\alpha(A) = M$, define the k -dimensional measure of M to be

$$\lambda_k(M) = \int_M dM = \int_A \sqrt{\det(D\alpha^T D\alpha)}.$$

This is a special case of the following more general definition.

Definition 10.4.4. Given a parametrized k -manifold $\alpha : U \rightarrow \mathbb{R}^n$, a measurable subset $A \subset U$ with $\alpha(A) = M$, and a function $f : M \rightarrow Y$, where $(Y, \|\cdot\|_Y)$ is a Banach space, if $f\sqrt{\det(D\alpha^\top D\alpha)}$ is integrable on U , define the integral

$$\int_M f \, dM = \int_A f \sqrt{\det(D\alpha^\top D\alpha)}.$$

Example 10.4.5. If $k = 1$, the term $\sqrt{\det(D\alpha^\top D\alpha)}$ in the previous definition is just $\|D\alpha\|_2 = \|\alpha'\|_2$, so that definition reduces to the formula for line integral of the function f . If $C = \alpha([a, b])$, we have

$$\int_C f \, dC = \int_{[a,b]} f \sqrt{\det(D\alpha^\top D\alpha)} = \int_a^b f \|\alpha'(t)\| \, dt = \int_C f \, ds.$$

Proposition 10.4.6. The value of the integral $\int_M f \, dM$ is independent of the choice of parametrization α .

Proof. The proof is Exercise 10.14. $\square$

Remark 10.4.7. For a surface S in $\mathbb{R}^3$, you may have seen surface integrals defined differently—not with $\sqrt{\det(D\alpha^\top D\alpha)}$, but rather as

$$\int_S f \, dS = \int_U f(\alpha) \|D_1\alpha \times D_2\alpha\| = \iint_U f(\alpha(u, v)) \|\alpha_u \times \alpha_v\| \, du \, dv. \quad (10.2)$$

This is equivalent to Definition 10.4.4, as can be seen by verifying that the area of the parallelogram defined by two vectors $\mathbf{v}_1, \mathbf{v}_2$ in $\mathbb{R}^3$ is also equal to the length $\|\mathbf{v}_1 \times \mathbf{v}_2\|$ of the cross product. For more on properties of cross products, see Appendix C.3.

10.4.3 Surface Integrals of Vector Fields

For the special case in which F is a vector field (that is, when $F : M \rightarrow \mathbb{R}^n$), we defined a special line integral by taking the inner product of F with the unit tangent vector $\mathbf{T}$. Although there is no uniquely determined tangent vector on a surface or higher-dimensional manifold, there is a uniquely determined normal vector on a surface S in $\mathbb{R}^3$. This partially motivates the following definition, which is also motivated physically by the fact that the following integral describes the *flux* across S (volume of fluid crossing the surface S per unit time) of a fluid moving with velocity F .

Definition 10.4.8. Given a parametrized surface $\alpha : U \rightarrow \mathbb{R}^3$, given a measurable subset $A \subset U$ with $S = \alpha(A)$, and given a function $F : S \rightarrow \mathbb{R}^3$, let $\mathbf{N} : U \rightarrow \mathbb{R}^3$ be the unit normal to α . If $(F \cdot \mathbf{N})\sqrt{\det(D\alpha^\top D\alpha)}$ is integrable on A , define the integral

$$\int_S F \cdot d\mathbf{S} = \int_S (F \cdot \mathbf{N}) dS = \int_A (F \cdot \mathbf{N}) \sqrt{\det(D\alpha^T D\alpha)}.$$

Remark 10.4.9. Combining the definition of the unit normal with the relation (10.2), we can also write the surface integral of a vector field as

$$\int_S F \cdot d\mathbf{S} = \int_A (F \cdot \mathbf{N}) \sqrt{\det(D\alpha^T D\alpha)} = \iint_A F(\alpha) \cdot (\alpha_u \times \alpha_v) du dv.$$

Proposition 10.4.10. *The surface integral of a vector-valued function is independent of the parametrization (but will change sign if the orientation is reversed).*

Proof. The proof is Exercise 10.15. $\square$

10.5 Green's Theorem

Green's theorem is a two-dimensional analogue of the fundamental theorem of calculus. Recall that the single-variable fundamental theorem of calculus says the integral over an interval of the derivative of a function can be computed from the value of the function on the boundary. Green's theorem says that the integral of a certain derivative of a function over a region in $\mathbb{R}^2$ can be computed from (an integral of) the function on the boundary of the surface.

Like the fundamental theorem of calculus, Green's theorem is a powerful tool both for proving theoretical results and for computing integrals that would be very difficult to compute any other way.

10.5.1 Jordan Curve Theorem

We begin with a definition and a theorem that seems obvious, but which is actually quite tricky to prove; it is called the Jordan curve theorem.

Definition 10.5.1. *A connected component of a subset $S \subset \mathbb{F}^n$ is a subset $T \subset S$ that is connected (see Definition 5.9.10) and is not contained in any other connected subset of S .*

Example 10.5.2.

- (i) The set $\mathbb{R}^1 \setminus \{0\}$ consists of two connected components, namely, the sets $(-\infty, 0)$ and $(0, \infty)$.
- (ii) The set $\mathbb{R}^n$ has only one connected component, namely, itself.
- (iii) The set $\{(x, y) \in \mathbb{R}^2 \mid xy \neq 0\}$ has four connected components, namely, the four quadrants of the plane.

Theorem 10.5.3 (Jordan Curve Theorem). *Let γ be a simple closed curve in $\mathbb{R}^2$ (see Definition 10.1.1). The complement $\mathbb{R}^2 \setminus \gamma$ consists of two connected*

components. One of these is bounded and one is unbounded. The curve γ is the topological boundary of each component.

Remark 10.5.4. We do not prove this theorem because it would take us far from our goals for this text. A careful proof is actually quite difficult.

Definition 10.5.5. We call the bounded component of the complement of the curve γ the interior of γ and the unbounded component of the complement the exterior of γ . If a point $\mathbf{z}_0$ lies in the interior of γ , we say that $\mathbf{z}_0$ is enclosed by γ or that it lies within γ .

Remark 10.5.6. Because $\mathbb{C}$ is homeomorphic to the plane $\mathbb{R}^2$ (see Definition 5.9.1), the Jordan curve theorem also applies to simple closed curves in $\mathbb{C}$.

Definition 10.5.7. We say that an open set $U \subset \mathbb{C}$ or in $\mathbb{R}^2$ is simply connected if for any simple closed curve γ that lies inside U , every point in the interior of γ is also contained in U . See Figure 10.5 for an example.

Nota Bene 10.5.8. Intuitively, U is simply connected if it contains no holes.

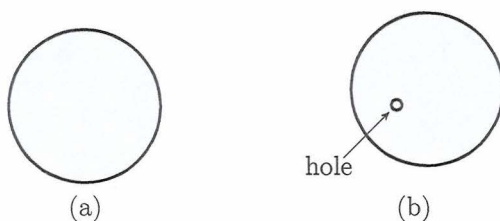


Figure 10.5. A region (a) in the plane that is simply connected, and one (b) that is not simply connected; see Definition 10.5.7.

Example 10.5.9.

- (i) The plane $\mathbb{R}^2$ is simply connected.
- (ii) The open ball $B(0, r) \subset \mathbb{R}^2$ is simply connected for every $r \in (0, \infty)$.

Unexample 10.5.10.

- (i) The annulus $\{(x, y) \in \mathbb{R}^2 \mid 0 < \|(x, y)\| < 1\}$ is not simply connected.
- (ii) The set $S = \mathbb{C} \setminus B(0, r)$ is not simply connected, because for any $\varepsilon > 0$ the circle $\{z \in \mathbb{C} \mid \|z\| = r + \varepsilon\}$ lies in S , and the origin lies in the interior of the circle, but the origin is not in S .

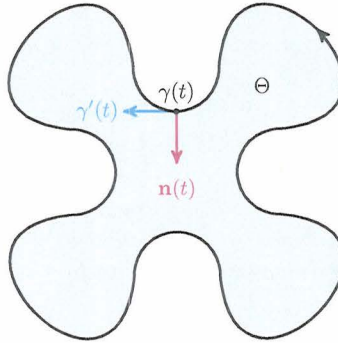


Figure 10.6. As described in Definition 10.5.12, a parametrized curve is positively oriented when the interior Θ lies to the left of the tangent vector $\gamma'(t)$ (blue). The left-pointing normal vector $\mathbf{n}(t)$ (red) points into Θ , and $-\mathbf{n}(t)$ points out of Θ .

Proposition 10.5.11. The interior of any simple closed curve in the plane is simply connected.

Proof. Let γ be a simple closed curve, and let $\mathbb{R}^2 \setminus \gamma$ be the disjoint union of connected components U and B , with U unbounded and B bounded. For any simple closed curve $\sigma \subset B$ we take $\mathbb{R}^2 \setminus \sigma = v \cup \beta$ with v unbounded and β bounded. Since U is connected and misses σ , we must have $U \subset v$ or $U \subset \beta$. Since U is unbounded, it cannot lie in β , so we have $U \subset v$, and hence $\beta \subset v^c \subset U^c = B \cup \gamma$. But $\gamma \cap \beta = \emptyset$ so $\beta \subset B$. $\square$

Definition 10.5.12. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed curve with interior $\Theta \subset \mathbb{R}^2$. Writing $\gamma = (x(t), y(t))$, we define the left-pointing normal vector $\mathbf{n}(t)$ at t to be $\mathbf{n}(t) = (-y'(t), x'(t))$.

We say that γ is positively oriented if for any $t \in [a, b]$ there is a $\delta > 0$ such that $\gamma(t) + h\mathbf{n}(t) \in \Theta$ whenever $0 < h < \delta$.

Remark 10.5.13. Said informally, γ has positive orientation if Θ always lies to the left of the tangent vector $\gamma'(t)$, or, equivalently, if γ is traversed in a counterclockwise direction. See Figure 10.6 for a depiction of this.

10.5.2 Green's Theorem

Definition 10.5.14. We say that a closed subset $\Delta \subset \mathbb{R}^2$ is an x -simple region if there is an interval $[a, b] \in \mathbb{R}$ and continuous⁴³ functions $f, g : [a, b] \rightarrow \mathbb{R}$ that are C^1 on (a, b) such that $\Delta = \{(x, y) \mid x \in [a, b], f(x) \leq y \leq g(x)\}$. Similarly, we say that Δ is a y -simple region if there is an interval $[a, b] \in \mathbb{R}$ and functions $f, g : [a, b] \rightarrow \mathbb{R}$ such that $\Delta = \{(x, y) \mid y \in [a, b], f(y) \leq x \leq g(y)\}$. See Figure 10.7 for examples. We say that Δ is a simple region if it is both x -simple and y -simple; see Figure 10.8

⁴³See Remark 10.1.2.

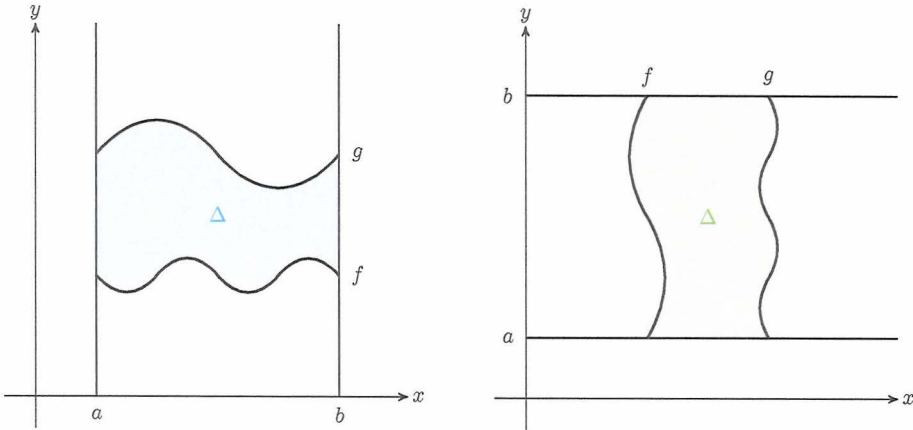


Figure 10.7. Examples of an x -simple region (left) and a y -simple region (right); see Definition 10.5.14.

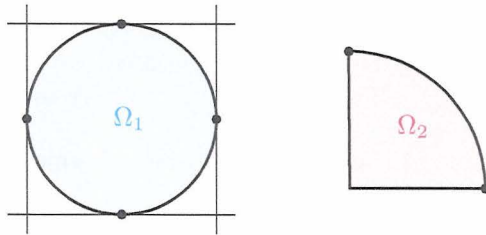


Figure 10.8. Examples of regions that can be classified as both x -simple and y -simple, as described in Definition 10.5.14.

Theorem 10.5.15 (Green's Theorem). Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a piecewise-smooth, positively oriented, simple closed curve with interior $\Omega \subset \mathbb{R}^2$ such that $\bar{\Omega} = \Omega \cup \gamma$ is the union of a finite number of simple regions $\Delta_1, \dots, \Delta_m$ for which the intersection $\Delta_i \cap \Delta_j$ has measure zero for every $i \neq j$.

If U is an open set containing $\bar{\Omega}$ and if $P, Q : U \rightarrow \mathbb{R}$ are C^1 , then we have

$$\int_{\bar{\Omega}} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \int_{\gamma} (P, Q) \cdot d\gamma = \int_{\gamma} P dx + Q dy.$$

Proof. It suffices to prove the theorem in the case that Ω is simple. We show that $\int_{\bar{\Omega}} \frac{\partial Q}{\partial x} = \int_{\gamma} Q dy$ in the case that Ω is x -simple. The proof that $-\int_{\bar{\Omega}} \frac{\partial P}{\partial y} = \int_{\gamma} P dx$ when Ω is y -simple is essentially the same.

If Ω is x -simple, we can write $\bar{\Omega} = \{(x, y) \mid x \in [a, b], f(x) \leq y \leq g(x)\}$. We may take γ to be the sum of four curves $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, traversed consecutively, where $\gamma_1(t), \gamma_3(t) : [a, b] \rightarrow \mathbb{R}^2$ are given by $\gamma_1(t) = (t, f(t))$ (the graph of f , traversed from left to right) and $\gamma_3(t) = (b + a - t, g(b + a - t))$ (the graph of g traversed from right to left), respectively, and $\gamma_2, \gamma_4 : [0, 1] \rightarrow \mathbb{R}^2$ are given by

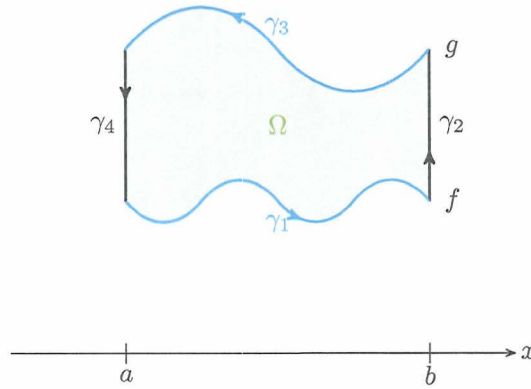


Figure 10.9. The x -simple region Ω used in the proof of Green's theorem (Theorem 10.5.15). Here γ_1 is the graph of f , traversed from left to right, while γ_3 is the graph of g , traversed from right to left. The path γ_2 is the right-hand vertical line, traversed from bottom to top, and γ_4 is the left-hand vertical line, traversed from top to bottom.

$\gamma_2(t) = (b, (1-t)f(b) + tg(b))$ (the right vertical line, traversed from bottom to top) and $\gamma_4(t) = (a, (1-t)g(a) + tf(a))$ (the left vertical line, traversed from top to bottom), respectively. See Figure 10.9 for a depiction of Ω and the four curves that make up γ .

By Fubini's theorem, the general Leibniz integral rule (8.10), and the fundamental theorem of calculus, we have

$$\begin{aligned}
 \int_{\Omega} \frac{\partial Q}{\partial x} &= \int_a^b \int_{f(x)}^{g(x)} \frac{\partial Q}{\partial x} dy dx \\
 &= \int_a^b \left[\frac{d}{dx} \left(\int_{f(x)}^{g(x)} Q(x, y) dy \right) + f'(x)Q(x, f(x)) - g'(x)Q(x, g(x)) \right] dx \\
 &= \int_{f(b)}^{g(b)} Q(b, y) dy - \int_{f(a)}^{g(a)} Q(a, y) dy \\
 &\quad + \int_a^b f'(x)Q(x, f(x)) dx - \int_a^b g'(x)Q(x, g(x)) dx \\
 &= \int_{\gamma_2} Q dy + \int_{\gamma_4} Q dy + \int_{\gamma_1} Q dy + \int_{\gamma_3} Q dy \\
 &= \int_{\gamma} Q dy. \quad \square
 \end{aligned}$$

Example 10.5.16. Let D be the upper half of the unit disk, so

$$D = \{(x, y) \in \mathbb{R}^2 \mid 0 < \|(x, y)\| \leq 1, y \geq 0\},$$

and let ∂D be the boundary curve, traversed once counterclockwise. To compute the line integral

$$\int_{\partial D} x^2 dx + 2xy dy$$

we may apply Green's theorem with $P(x, y) = x^2$ and $Q(x, y) = 2xy$:

$$\int_{\partial D} x^2 dx + 2xy dy = \iint_D \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \iint_D 2y dx dy$$

Converting to polar coordinates gives

$$\iint_D 2y dx dy = \int_0^\pi \int_0^1 2r \sin(\theta) r dr d\theta = 4/3.$$

Remark 10.5.17. Although the statement of the theorem only mentions simple closed curves (and hence simply connected regions), we can easily extend it to more general cases. For example, if Ω is the region shown in Figure 10.10, we can cut Ω into two pieces, $\Omega_1 \cup \Omega_2$, each of which is bounded by a simple closed piecewise-smooth curve.

Since each of the new cuts is traversed twice, once in each direction, the contribution of each cut to the line integral is zero. Also each cut has measure zero in the plane, so it also contributes zero to the integral $\iint_\Omega \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}$, so we have

$$\begin{aligned} \iint_\Omega \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} &= \sum_{i=1}^2 \iint_{\Omega_i} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \\ &= \sum_{i=1}^2 \int_{\gamma_i} (P, Q) \cdot d\gamma \\ &= \int_\gamma (P, Q) \cdot d\gamma + \int_\tau (P, Q) \cdot d\gamma. \end{aligned}$$

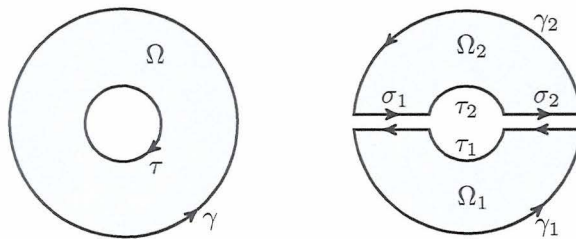


Figure 10.10. Two cross cuts σ_1 and σ_2 subdivide the annulus Ω into two simply connected regions, Ω_1 and Ω_2 . The boundary of Ω_1 is $\gamma_1 - \sigma_2 + \tau_1 - \sigma_1$ and the boundary of Ω_2 is $\gamma_2 + \sigma_1 + \tau_2 + \sigma_2$. When integrating along the boundary of both Ω_1 and Ω_2 , the contributions from $\pm\sigma_1$ and $\pm\sigma_2$ cancel, giving the same result as integrating along the boundary of Ω , that is, integrating over $\gamma + \tau$. See Remark 10.5.17 for details and Example 10.5.18 for a specific case.

Example 10.5.18. Consider the annulus $\Omega = \{(x, y) \in \mathbb{R}^2 \mid 1 \leq \|(x, y)\| \leq 2\}$. Let τ be the inner boundary circle oriented clockwise, and let γ be the outer boundary circle, oriented counterclockwise. To evaluate the integral

$$\int_{\tau \cup \gamma} y^3 dx - x^3 dy$$

cut the annulus along the x -axis and add new paths $\sigma_1 : [1, 2] \rightarrow \mathbb{R}^2$ and $\sigma_2 : [1, 2] \rightarrow \mathbb{R}^2$, given by $\sigma_1(t) = (t, 0)$ and $\sigma_2(t) = (-2 + t, 0)$.

Green's theorem applies to each half of the annulus to give

$$\begin{aligned} \int_{\tau + \gamma} y^3 dx - x^3 dy &= \left(\int_{\tau_1 + \gamma_1 - \sigma_1 - \sigma_2} y^3 dx - x^3 dy \right) \\ &\quad + \left(\int_{\tau_2 + \gamma_2 + \sigma_1 + \sigma_2} y^3 dx - x^3 dy \right) \\ &= \int_{\Omega_1} -3(x^2 + y^2) dx dy + \int_{\Omega_2} -3(x^2 + y^2) dx dy \\ &= \int_{\Omega} -3(x^2 + y^2) dx dy \\ &= \int_0^{2\pi} \int_1^2 (-3r^2) r dr d\theta = -45\pi/2. \end{aligned}$$

Remark 10.5.19. Green's theorem, when combined with the chain rule, can be applied to surfaces in $\mathbb{R}^3$ to give a special case of *Stokes' theorem*. Assume γ is a simple closed curve in the plane with interior Ω , and $\alpha : U \rightarrow \mathbb{R}^3$ is a C^1 surface with $\overline{\Omega} \subset U$. Let $S = \alpha(\overline{\Omega})$ with boundary C (parametrized by $\alpha(\gamma) = \sigma$).

For $F : S \rightarrow \mathbb{R}^3$ with $F = (P, Q, R)$, the *curl* of F is

$$\text{curl}(F) = \nabla \times F = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right). \quad (10.3)$$

The curl operator describes the infinitesimal rotation of F at each point. Stokes' theorem states that

$$\int_C F \cdot d\sigma = \int_S \text{curl}(F) \cdot d\mathbf{S}.$$

This theorem follows from a tedious but straightforward application of Green's theorem and the chain rule.

Vista 10.5.20. Both Stokes' and Green's theorems are special cases of a great generalization of the fundamental theorem of calculus. Roughly speaking, when integrating over a parametrized manifold M , there is a natural way to differentiate integrands called *exterior differentiation* that turns a k -dimensional integrand η into a $k + 1$ -dimensional integrand $d\eta$.

The generalized fundamental theorem of calculus can be written as

$$\int_{\partial M} \eta = \int_M d\eta, \quad (10.4)$$

where ∂M is the boundary of M . In the case of $k = 1$, if $\eta = P dx + Q dy$, then $d\eta = \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dx dy$, which agrees with Green's theorem. In the case of $k = 2$, if $\eta = F \cdot d\mathbf{S}$, then $d\eta = \text{curl}(F) \cdot d\mathbf{S}$, so the theorem reduces to Stokes' theorem.

The case where $k = 0$ is just the usual fundamental theorem of calculus, if we make the following fairly natural definitions. Let the zero-dimensional integral $\int_{\mathbf{p}} f$ of a function f at a point $\mathbf{p} \in \mathbb{R}$ be $f(\mathbf{p})$, let the integral at the negatively oriented point $-\mathbf{p}$ be $\int_{-\mathbf{p}} f = -f(\mathbf{p})$, and let $\partial[a, b]$ be the sum of the oriented points $+b$ and $-a$. Finally, for any C^1 function $f : [a, b] \rightarrow \mathbb{R}$ let $df = f'(t)dt$. Putting these together, we have

$$\int_{\partial[a,b]} f = f(b) - f(a) = \int_a^b f'(t)dt = \int_{[a,b]} df.$$

For more on this generalization of the fundamental theorem of calculus, see the references listed in the notes for this chapter.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

10.1. Let H be the curve with parametrization $\sigma : [0, 2\pi] \rightarrow \mathbb{R}^3$ given by $\sigma(t) = (r \cos(t), r \sin(t), ct)$ for some positive constants $r, c \in \mathbb{R}$. Find the arclength of H .

10.2. Prove Proposition 10.1.15.

10.3. Consider a wheel of radius 1 rolling along the x -axis in $\mathbb{R}^2$ without slipping. Fix a point on the outside of the wheel, and let C be the curve swept out by the point.

- (i) Find a parametrization $\sigma : \mathbb{R} \rightarrow \mathbb{R}^2$ of this curve.
- (ii) Find the distance traveled by the point when the wheel makes one full rotation.

10.4. Let $X = M_2(\mathbb{R})$ be the Banach space of 2×2 real matrices with the 1-norm, and let C be the curve with parametrization $\sigma : [0, 1] \rightarrow C$, given by

$$\sigma(t) = \begin{bmatrix} t + e^t & 1 \\ 5 & e^{-t} \end{bmatrix}.$$

For any $t \in (0, 1)$ find the unit tangent to C . Find the arclength of C .

10.5. Let X , C , and σ be as in Exercise 10.4. Let $f = \det : X \rightarrow \mathbb{R}$ be the determinant function. Evaluate the integral $\int_C f \, ds$.

10.6. Compute the line integral

$$\int_C y(z+1) \, dx + xz \, dy + xy \, dz,$$

where C is parametrized by $\sigma(\theta) = (\cos(\theta), \sin(\theta), \sin^3(\theta) + \cos^3(\theta))$, with $\theta \in [0, 2\pi]$.

10.7. Compute the line integral

$$\int_C \frac{x \, dy - y \, dx}{x^2 + y^2},$$

where C is a circle of radius R around the origin beginning and ending at $(R, 0)$ and proceeding counterclockwise.

- (i) How does changing R change the answer?
- (ii) How do the results change if the orientation is reversed?
- (iii) How does the result change if C is a different simple closed curve around the origin (not necessarily a circle)? Hint: Note that F is conservative (see Example 10.2.10) in certain parts of the plane—if $u = -\text{Arctan}(x/y)$, then we have $F = Du$ everywhere but on the line $y = 0$, and if we set $v = \text{Arctan}(x/y)$, we have $F = Dv$ everywhere but on the line $x = 0$.

10.8. Prove Proposition 10.2.2.

10.9. Prove Proposition 10.2.8.

10.10. Let S be the surface

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid 9xz - 7y^2 = 0\}.$$

- (i) Show that the function $\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, given by $\alpha(t, u) = (t^2, 3ut, 7u^2)$, is C^1 on all of $\mathbb{R}^2$ and satisfies $\alpha(t, u) \in S$ for all $(t, u) \in \mathbb{R}^2$.
- (ii) Show that α is not injective.
- (iii) Find all the points of the domain where $D\alpha$ is not injective.

- (iv) Find an open subset U of $\mathbb{R}^2$ such that the point $\mathbf{p} = (1, 3, 7) \in S$ lies in $\alpha(U)$ and such that α restricted to U is a parametrized 2-manifold.
 - (v) Give a basis for the tangent space $T_{\mathbf{p}}S$ of S at $\mathbf{p}$.
 - (vi) Give the unit normal $\mathbf{N}$ of S at $\mathbf{p}$ with the parametrization α .
- 10.11. Let $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be given by $(t, u) \mapsto (u, t)$. Let S , $\mathbf{p}$, U , and α be as in the previous problem.
- (i) Find a basis of $T_{\mathbf{p}}S$ using the parametrization $\alpha \circ \phi$. Prove that the span of this basis is the same as the span of the basis computed with the parametrization α .
 - (ii) Compute the unit normal at $\mathbf{p}$ for the parametrization $\alpha \circ \phi$. Verify that this is the same as the unit normal computed for α , except that the sign has changed.
- 10.12. For each of the parametrizations α , β , and σ of Examples 10.3.5 and 10.3.7, compute the unit normal at the point $(1/2, 0, \sqrt{3}/2)$. Verify that it is the same for α and β , but has the opposite sign for σ .
- 10.13. Describe the tangent space $T_{\mathbf{p}}M$ at each point of each of the following manifolds:
- (i) A vector subspace of $\mathbb{R}^n$.
 - (ii) The graph of a smooth function $f : \mathbb{R}^k \rightarrow \mathbb{R}$.
 - (iii) The cylinder parametrized by $u, v \mapsto (u, \cos(v), \sin(v))$.

10.14.* Prove Proposition 10.4.6.

10.15.* Prove Proposition 10.4.10.

10.16.* Prove that the surface area of a sphere $S \subset \mathbb{R}^3$ of radius r is $4\pi r^2$ by parametrizing half of the sphere and computing the integral $\int_S dS$.

10.17.* Set up an integral to compute the n -dimensional “surface area” of the n -sphere of radius r in $\mathbb{R}^{n+1}$. You need not evaluate the integral.

10.18.* Let M be the surface in $\mathbb{R}^4$ defined by the equation $z_2 = z_1^2$ in $\mathbb{C}^2$ (thinking of $\mathbb{C}^2$ as $\mathbb{R}^4$). Compute the surface area of the part of M satisfying $|z_1| \leq 1$. Hint: Write the surface in polar coordinates as $\alpha(r, \theta) = (r \cos(\theta), r \sin(\theta), r^2 \cos(2\theta), r^2 \sin(2\theta))$, with $r \in [0, 1]$.

10.19. Let $C \subset \mathbb{R}^2$ be the simple closed curve described in polar coordinates by $r = \cos(2\theta)$, where $\theta \in [-\pi/4, \pi/4]$, traversed in a counterclockwise direction. Compute the integral

$$\int_C 3y \, dx + x \, dy.$$

10.20. Evaluate the integral $\int_C (x - y^3) \, dx + x^3 \, dy$, where C is the unit circle $x^2 + y^2 = 1$ traversed once counterclockwise.

10.21. Use Green's theorem to give another proof of the result in Example 10.2.10 that if $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ is C^2 and $F = Du$, then for any simple closed curve C in the plane satisfying the hypothesis of Green's theorem and parametrized by $\sigma : [a, b] \rightarrow \mathbb{R}^2$ we have

$$\int_C F \cdot d\sigma = 0.$$

- 10.22. Let $D = \{z \in \mathbb{C} : |z| \leq 1\}$ be the unit disk, and let C be the boundary circle $\{(x, y) \mid x^2 + y^2 = 1\}$, traversed once counterclockwise. Let

$$F(x, y) = (P(x, y), Q(x, y)) = \begin{cases} (-y, x)/(x^2 + y^2) & \text{if } (x, y) \neq (0, 0), \\ (0, 0) & \text{otherwise.} \end{cases}$$

Evaluate both of the integrals

$$\int_D \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \quad \text{and} \quad \int_C P dx + Q dy$$

Explain why your results do not contradict Green's theorem.

- 10.23.* Let γ be a simple closed curve in the plane, and let Ω be its interior, with $\gamma \cup \Omega$ contained in an open set $U \subset \mathbb{R}^2$. Assume $\gamma(t) = (x(t), y(t))$ is positively oriented, and let $d\mathbf{n} = (y'(t), -x'(t)) dt$ be the outward-pointing normal. For any C^1 vector field $F : U \rightarrow \mathbb{R}^2$, let $\nabla \cdot F = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y}$. Prove that

$$\int_{\Omega} \nabla \cdot F = \int_{\gamma} F \cdot d\mathbf{n}.$$

- 10.24.* Using the same assumptions as the previous exercise, let $\nabla \times F = \text{curl}(F)$ be as in (10.3), and let $\mathbf{e}_3 = (0, 0, 1)$ be the third standard basis vector. Prove that

$$\int_{\Omega} (\nabla \times F) \cdot \mathbf{e}_3 = \int_{\gamma} F \cdot \mathbf{T} ds.$$

Notes

The interested reader can find proofs of the Jordan curve theorem in [Bro00, Hal07a, Hal07b, Mae84]. For more on generalizations of the cross product, the exterior algebra, differential forms, and the generalized fundamental theorem of calculus, including its relation to Green's and Stokes' theorems, we recommend the books [HH99] and [Spi65].

11

Complex Analysis

Between two truths of the real domain, the easiest and shortest path quite often passes through the complex domain.

—Paul Painlevé

In this chapter we discuss properties of differentiable functions of a single complex variable. You might think that we have completely treated differentiable functions in Chapter 6 and that complex differentiable functions should just be a special case of that more general setting, but some remarkable and important things happen in this special case that do not happen in the general setting. Differentiability in this setting is a very strong condition that allows us to prove many results about complex functions that we use in Chapters 12 and 13 to prove some powerful results about linear operators.

Throughout this chapter, let $(X, \|\cdot\|)$ be a Banach space over $\mathbb{C}$.

11.1 Holomorphic Functions

11.1.1 Differentiation on $\mathbb{C}$

The complex plane $\mathbb{C} = \{x + iy \mid x, y \in \mathbb{R}\}$ is naturally a normed $\mathbb{R}$ -vector space, with scalar multiplication $\alpha z = \alpha(x + iy) = (\alpha x + i\alpha y)$, with vector addition the same as complex addition, $(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$, and with norm $|(x + iy)| = \sqrt{x^2 + y^2}$. There is an obvious bijective map from $\mathbb{C}$ to $\mathbb{R}^2$, given by sending $z = x + iy \in \mathbb{C}$ to $(x, y) \in \mathbb{R}^2$, and this map is an isomorphism of normed $\mathbb{R}$ -vector spaces, where we use the 2-norm on $\mathbb{R}^2$. This implies that $\mathbb{C}$ and $\mathbb{R}^2$ have the same topology. Throughout the rest of the book we often use this fact and the associated isomorphism $x + iy \mapsto (x, y)$ without further mention.

Let $U \subset \mathbb{C}$ be an open set in $\mathbb{C}$. The fact that we can think of $\mathbb{C}$ in two ways—as both a two-dimensional real vector space and as a one-dimensional complex vector space, leads to two different notions of the total derivative of a function $f: U \rightarrow X$. First, if we consider $\mathbb{C}$ as a two-dimensional real vector space and X as a Banach space over $\mathbb{R}$, the total derivative at each point is a linear transformation from $\mathbb{R}^2$ to X . So, for example, if the target X is the real vector space $\mathbb{C} \cong \mathbb{R}^2$, then the

total real derivative of f is represented by the 2×2 Jacobian matrix, provided the partial derivatives exist and are continuous.

But if we consider the domain $\mathbb{C}$ as a one-dimensional complex vector space, then the total derivative at each point is a bounded $\mathbb{C}$ -linear transformation from $\mathbb{C}$ to X . In this case the total complex derivative at a point z_0 is given by

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} \in X.$$

More precisely, for any $w \in \mathbb{C}$ the $\mathbb{C}$ -linear map $Df(z_0) : \mathbb{C} \rightarrow X$ is given by $Df(z_0)(w) = wf'(z_0)$, where $wf'(z_0)$ means multiplication of the vector $f'(z_0) \in X$ by the complex scalar w (see Remark 6.2.7). When we say that a complex function is differentiable, we always mean it is differentiable in this second sense, unless otherwise specified.

Definition 11.1.1. Let $f : U \rightarrow X$, where $U \subset \mathbb{C}$ is open. We say that the function f is holomorphic on U if it is differentiable (see Definition 6.3.1) at every point of U , considering U as an open subset of the one-dimensional complex vector space $\mathbb{C}$; that is, f is holomorphic on U if the limit

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} \quad (11.1)$$

exists for every $z_0 \in U$, where the limits are taken with z and h in $\mathbb{C}$ (not just in $\mathbb{R}$). If U is not specified, then saying f is holomorphic means that f is holomorphic on all of $\mathbb{C}$. If f is holomorphic on all of $\mathbb{C}$, it is sometimes said to be entire.

Example 11.1.2. We show that $f(z) = z^2$ is a holomorphic function. Let $h = re^{i\theta}$. We compute

$$\begin{aligned} f'(z_0) &= \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h} \\ &= \lim_{r \rightarrow 0} \frac{(z_0 + re^{i\theta})^2 - z_0^2}{re^{i\theta}} \\ &= \lim_{r \rightarrow 0} \frac{2z_0 re^{i\theta} + (re^{i\theta})^2}{re^{i\theta}} \\ &= \lim_{r \rightarrow 0} 2z_0 + re^{i\theta} = 2z_0. \end{aligned}$$

Thus $f(z) = z^2$ is holomorphic with derivative $2z_0$ at z_0 .

11.1.2 The Cauchy–Riemann Equations

The condition of being holomorphic is a much stronger condition than being differentiable as a function of two real variables. If f is holomorphic, then it is also differentiable as a map from $\mathbb{R}^2$; but the converse is not true.

The reason that real differentiability does not imply complex differentiability is that in the complex case the linear transformation $Df(z_0) = f'(z_0)$ is given by complex scalar multiplication $w \mapsto wf'(z_0)$, whereas most real linear transformations from $\mathbb{R}^2$ to X are *not* given by complex scalar multiplication. Of course any $\mathbb{C}$ -linear map is $\mathbb{R}$ -linear, but the converse is certainly not true, as the next example shows.

Unexample 11.1.3. Consider the function $f(z) = \bar{z}$. We show that this is a real differentiable function of two variables but that it is not holomorphic. Let $z = x + iy$, so that $f(z) = f(x, y) = x - iy$. Considered as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$, this is a differentiable function, and the total derivative is

$$Df(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

For this function to be holomorphic, the limit (11.1) must exist regardless of the direction in which z approaches z_0 . Suppose $z_0 = 0$, and we let $h = t$ where $t \in \mathbb{R}$. If the derivative $f'(z_0) = \lim_{h \rightarrow 0} (f(z_0 + h) - f(z_0))/h$ exists, then it must be equal to

$$\lim_{t \rightarrow 0} \frac{f(t) - f(0)}{t} = \lim_{t \rightarrow 0} \frac{t - 0}{t} = 1.$$

On the other hand, if $h = it$ with $t \in \mathbb{R}$, then whenever the derivative $f'(z_0) = \lim_{h \rightarrow 0} (f(z_0 + h) - f(z_0))/h$ exists, it must be equal to

$$\lim_{t \rightarrow 0} \frac{f(it) - 0}{it} = \lim_{t \rightarrow 0} \frac{-it - 0}{it} = -1.$$

Since these two real limits are not equal, the complex limit does not exist, and the function cannot be holomorphic.

Any $\mathbb{C}$ -linear map from $\mathbb{C}$ to X is determined by a choice of $\mathbf{b} \in X$, and the map is given by

$$x + iy \mapsto (x + iy)\mathbf{b} = x\mathbf{b} + iy\mathbf{b}.$$

But an $\mathbb{R}$ -linear map from $\mathbb{R}^2$ to X is determined by a choice of two elements, $\mathbf{b}_1, \mathbf{b}_2 \in X$, and the map is given by

$$x + iy \mapsto x\mathbf{b}_1 + y\mathbf{b}_2.$$

Thus, the $\mathbb{R}$ -linear map defined by $\mathbf{b}_1, \mathbf{b}_2$ is $\mathbb{C}$ -linear if and only if $\mathbf{b}_2 = i\mathbf{b}_1$.

In terms of derivatives of a function $f: \mathbb{C} \rightarrow X$, this means that the total real derivative $[\partial f/\partial x \quad \partial f/\partial y]$ can only be $\mathbb{C}$ -linear, and hence can only define a complex derivative $f'(z) = \partial f/\partial x + i\partial f/\partial y$, if it satisfies

$$\frac{\partial f}{\partial y} = i \frac{\partial f}{\partial x}. \quad (11.2)$$

This important equation is called the *Cauchy–Riemann* equation. In the special case that $X = \mathbb{C}$, we can write $f(x, y) = u(x, y) + iv(x, y)$, so that $\partial f/\partial x = \partial u/\partial x + i\partial v/\partial x$ and $\partial f/\partial y = \partial u/\partial y + i\partial v/\partial y$. Expanding (11.2) gives

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \quad (11.3)$$

which, together, are also often called the *Cauchy–Riemann equations*. These equations are an important way to determine whether a function is holomorphic.

Theorem 11.1.4 (Cauchy–Riemann). *Let $U \subset \mathbb{C}$ be an open set, and let $f: U \rightarrow X$. If f is holomorphic on U , then it is also real differentiable on U , and the partials $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist. Moreover, (11.2) holds for all $(x, y) \in U \subset \mathbb{R}^2$. The converse also holds: if f is real differentiable and the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, exist, are continuous, and satisfy (11.2) on $U \subset \mathbb{C}$, then f is holomorphic on U .*

Proof. Assume that $f'(z_0)$ exists for $z_0 \in U$. Let $z_0 = x_0 + iy_0$ and $z = x + iy_0$ where $(x_0, y_0) \in \mathbb{R}^2$ and $x \in \mathbb{R}$. As $x \rightarrow x_0$ we have

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{x \rightarrow x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0} = \left. \frac{\partial f}{\partial x} \right|_{z=z_0}. \quad (11.4)$$

Similarly, if $z = x_0 + iy$, then as $y \rightarrow y_0$, we have

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{y \rightarrow y_0} \frac{f(x_0, y) - f(x_0, y_0)}{i(y - y_0)} = -i \left. \frac{\partial f}{\partial y} \right|_{z=z_0}.$$

Since $f'(z_0)$ exists, these two real limits must exist and be equal, which gives the Cauchy–Riemann equation (11.2). If $X = \mathbb{C}$, we can write $f(x, y) = u(x, y) + iv(x, y)$, expand (11.2), and match the real parts to real parts, and imaginary parts to imaginary parts to get (11.3).

Conversely, if the real partials of f exist and are continuous, then the total real derivative exists and is given by $Df(x, y) = [\frac{\partial f}{\partial x} \frac{\partial f}{\partial y}]$. By Lemma 6.4.1 (the alternative characterization of the Fréchet derivative) applied to the function f on $U \subset \mathbb{R}^2$ at any $z_0 = x_0 + iy_0 \in U$, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $(a, b) \in \mathbb{R}^2$ satisfies $\|(a, b)\|_2 < \delta$, then we have

$$\|f(x_0 + a, y_0 + b) - f(x_0, y_0) - Df(x_0, y_0)(a, b)\| \leq \varepsilon \|(a, b)\|_2.$$

Let $g(z) = \frac{\partial f}{\partial x} \in X$. If (11.2) holds, then we have already seen that $Df(x_0, y_0)$ acts on elements (a, b) complex linearly—by sending (a, b) to $(a + bi)g(z_0) \in X$, thought of as an element of $\mathcal{B}(\mathbb{C}; X) \cong X$. But this shows that if $h = a + bi$ has $|h| < \delta$, then

$$\|f(z_0 + h) - f(z_0) - gh\| \leq |h|\varepsilon,$$

which implies by Lemma 6.4.1 that $g = f'(z)$. $\square$

Corollary 11.1.5. *If $U \subset \mathbb{C}$ is open, and if all of the partial derivatives of $u(x + iy) : U \rightarrow \mathbb{R}$ and $v(x + iy) : U \rightarrow \mathbb{R}$ exist on U , are continuous on U , and*

satisfy the Cauchy–Riemann equations (11.3), then the function $f: \mathbb{C} \rightarrow \mathbb{C}$ given by $f(x + iy) = u(x, y) + iv(x, y)$ is holomorphic on U .

Proof. This follows immediately from the previous theorem and the fact that continuous partials imply differentiability (see Theorem 6.2.14). $\square$

Unexample 11.1.6. The function $f(z) = \frac{z}{|z|^2}$ is not holomorphic on any open set. To see this, note that $f = \frac{x+iy}{x^2+y^2}$, so $u(x, y) = \frac{x}{x^2+y^2}$ and $v(x, y) = \frac{y}{x^2+y^2}$, and

$$\frac{\partial u}{\partial x} = \frac{y^2 - x^2}{(x^2 + y^2)^2} \quad \text{but} \quad \frac{\partial v}{\partial y} = \frac{x^2 - y^2}{(x^2 + y^2)^2}.$$

These are only equal if $x^2 = y^2$. Similarly

$$\frac{\partial u}{\partial y} = \frac{-2xy}{(x^2 + y^2)^2} \quad \text{and} \quad -\frac{\partial v}{\partial x} = \frac{2xy}{(x^2 + y^2)^2}.$$

These are equal only if $xy = 0$. The only point where $x^2 = y^2$ and $xy = 0$ is $x = y = 0$, so the origin is the only point of $\mathbb{C}$ where the Cauchy–Riemann equations could hold; but f is not even defined at the origin. Since the Cauchy–Riemann equations do not hold on any open set, f is not holomorphic on any open set.

The following, somewhat surprising result is a consequence of the Cauchy–Riemann equations and gives an example of how special holomorphic functions are.

Proposition 11.1.7. *Let f be $\mathbb{C}$ -valued and holomorphic on a path-connected open set U . If $\bar{f}$ is holomorphic on U or if $|f|$ is constant on U , then f is constant on U .*

Proof. The proof is Exercise 11.5. $\square$

11.2 Properties and Examples

In this section we survey some of the basic properties and examples of holomorphic functions. Probably the most important holomorphic functions are those that are complex valued—that is, functions from $U \subset \mathbb{C}$ to $\mathbb{C}$. But we also need to think about functions that take values in more general Banach spaces over $\mathbb{C}$. This is especially useful in Chapter 12, where we use complex-analytic techniques to develop results about linear operators.

After describing some of the elementary properties of holomorphic functions, we prove that every function that can be written as a convergent power series (these are called *analytic functions*) is holomorphic. We conclude the section with a proof of *Euler’s formula*.

11.2.1 Basic Properties

Remark 11.2.1. Holomorphic functions are continuous because they are differentiable (see Corollary 6.3.8).

Theorem 11.2.2. Let f, g be holomorphic, complex-valued functions on the open set $U \subset \mathbb{C}$. The following hold:

- (i) $af + bg$ is holomorphic on U and $(af + bg)' = af' + bg'$, where $a, b \in \mathbb{C}$.
- (ii) fg is holomorphic on U and $(fg)' = f'g + fg'$.
- (iii) If $g(z) \neq 0$ on U , then f/g is holomorphic on U and

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

- (iv) Polynomials are holomorphic on all of $\mathbb{C}$ (they are entire) and

$$(a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0)' = na_n z^{n-1} + (n-1)a_{n-1} z^{n-2} + \cdots + a_1.$$

- (v) A rational function

$$\frac{a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0}{b_m z^m + b_{m-1} z^{m-1} + \cdots + b_1 z + b_0}$$

is holomorphic on all of $\mathbb{C}$ except at the roots of the denominator.

Proof. The proof is Exercise 11.6. $\square$

Theorem 11.2.3 (Chain Rule). Let U, V be open sets in $\mathbb{C}$. If $f: U \rightarrow \mathbb{C}$ and $g: V \rightarrow X$ are holomorphic on U and V , respectively, with $f(U) \subset V$, then the map $h: U \rightarrow X$ given by $h(z) = g(f(z))$ is holomorphic on U and satisfies $h'(z) = g'(f(z))f'(z)$ for all $z \in U$.

Proof. This follows from Theorem 6.4.7. $\square$

Proposition 11.2.4. Let $U \subset \mathbb{C}$ be open and path connected. If $f: U \rightarrow X$ is holomorphic on U and $f'(z) = 0$ for all $z \in U$, then f is constant on U .

Proof. For any $z_1, z_2 \in U$, let $g: [0, 1] \rightarrow U$ be a smooth path satisfying $g(0) = z_1$ and $g(1) = z_2$. From the fundamental theorem of calculus (6.16), we have

$$f(z_2) - f(z_1) = f(g(1)) - f(g(0)) = \int_0^1 f'(g(t))g'(t) dt = 0.$$

Thus, we have $f(z_1) = f(z_2)$. Since the choice of z_1 and z_2 is arbitrary, the function f is constant. $\square$

11.2.2 Convergent Power Series Are Holomorphic

In this section we consider some basic properties of convergent power series over the complex numbers. Any function that can be written as a convergent power series is called *analytic*. The main result of this section is the fact (Theorem 11.2.8) that all analytic functions are holomorphic.

One of the most important theorems of complex analysis is the converse to Theorem 11.2.8, that is, if a function f is holomorphic in a neighborhood of a point z_0 , then f is analytic in some neighborhood of z_0 . This theorem is a consequence of the famous and powerful *Cauchy integral formula*, which we discuss in Section 11.4.

For every positive constant r and every $z_0 \in \mathbb{C}$, a polynomial of the form $f_n(z) = \sum_{k=0}^n a_k(z - z_0)^k$, with each $a_k \in X$, defines a function in the Banach space $(L^\infty(\overline{B(z_0, r)}; X), \|\cdot\|_{L^\infty})$. The discussion in Section 5.6.3 about convergence of series in a Banach space applies to series of the form $\sum_{k=0}^\infty a_k(z - z_0)^k$, and convergence with respect to the norm $\|\cdot\|_{L^\infty}$ is called uniform convergence. Whenever we talk about convergence of series, we mean convergence in this Banach space, unless explicitly stated otherwise. We also continue to use the definition of uniform convergence on an open set U to mean uniform convergence on every compact subset of U .

We begin with a useful lemma about convergence of power series in $\mathbb{C}$.

Lemma 11.2.5 (Abel–Weierstrass Lemma). *Let $(a_k)_{k \in \mathbb{N}} \subset X$ be a sequence. If there exists an $R > 0$ and $M \in \mathbb{R}$ such that*

$$\|a_n\|R^n \leq M$$

for every $n \in \mathbb{N}$, then for any positive $r < R$ the series $\sum_{k=0}^\infty a_k(z - z_0)^k$ and the series $\sum_{k=0}^\infty ka_k(z - z_0)^{k-1}$ both converge uniformly and absolutely on the closed ball $\overline{B(z_0, r)} \subset \mathbb{C}$.

Proof. Let D denote the closed ball $\overline{B(z_0, r)}$ and set $\rho = r/R$. For every $z \in D$ we have

$$\|a_k(z - z_0)^k\| \leq \|a_k\|r^k = \|a_k\|R^k\rho^k \leq M\rho^k.$$

Thus, on D we have

$$\sum_{k=0}^\infty \|a_k(z - z_0)^k\|_{L^\infty} \leq \sum_{k=0}^\infty M\rho^k.$$

Similarly, for every $z \in D$ we have

$$\|ka_k(z - z_0)^{k-1}\| \leq k\|a_k\|r^{k-1} = k\|a_k\|R^{k-1}\rho^{k-1} = k\|a_k\|R^k\frac{\rho^{k-1}}{R} \leq kM\frac{\rho^{k-1}}{R}.$$

Thus, on D we have

$$\sum_{k \in \mathbb{N}} \|ka_k(z - z_0)^{k-1}\|_{L^\infty} \leq \sum_{k \in \mathbb{N}} k\frac{M}{R}\rho^{k-1}.$$

The series $M \sum_{k=0}^\infty \rho^k$ and the series $M/R \sum_{k=0}^\infty k\rho^{k-1}$ both converge absolutely by the ratio test because $\rho < 1$. Hence $\sum_{k=0}^\infty a_k(z - z_0)^k$ and $\sum_{k \in \mathbb{N}} ka_k(z - z_0)^{k-1}$

both converge absolutely on every $\overline{B(z_0, r)} \subset B(z_0, R)$, and therefore they both converge uniformly on $B(z_0, R)$. $\square$

The following corollary is immediate.

Corollary 11.2.6. *If a series $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ does not converge at some point z_1 , then the series does not converge at any point $z \in \mathbb{C}$ with $|z - z_0| > |z_1 - z_0|$.*

Definition 11.2.7. *Given a power series $\sum_{k=0}^{\infty} a_k(z - z_0)^k$, the largest real number R such that the series converges uniformly on compact subsets in $B(z_0, R)$ is called the radius of convergence of the series. If the series converges for every $R > 0$, then we say the radius of convergence is ∞ .*

Theorem 11.2.8. *Given any power series $f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k$ that is uniformly convergent on compact subsets in an open ball $B(z_0, R)$, the function f is holomorphic on $B(z_0, R)$. The power series $g(z) = \sum_{k=1}^{\infty} k a_k(z - z_0)^{k-1}$ also converges uniformly and absolutely on compact subsets in $B(z_0, R)$, and $g(z)$ is the derivative of $f(z)$ on $B(z_0, R)$.*

Proof. By definition, the sequence of functions $f_n(z) = \sum_{k=0}^n a_k(z - z_0)^k$ converges uniformly on compact subsets to $f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k$. Similarly, the sequence of derivatives is $f'_n(z) = \sum_{k=1}^n k a_k(z - z_0)^{k-1}$. Since $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ converges on $\overline{B(z_0, r)}$ for any $0 < r < R$, the individual terms $a_k(z - z_0)^k \rightarrow 0$ for every $z \in \overline{B(z_0, r)}$, and there is a constant $M > 0$ such that $M \geq \|a_k\| r^k$ for every k .

Therefore, by the Abel-Weierstrass lemma, the series $\sum_{k \in \mathbb{N}} k a_k(z - z_0)^{k-1}$ converges absolutely to a function g on $\overline{B(z_0, r - \delta)}$ for every $\delta < r$. But since $r < R$ and $\delta < r$ are arbitrary, the series converges absolutely on every closed ball in $B(z_0, R)$, as required.

Finally, observe that the function g is continuous on $B(z_0, R)$, since it is the uniform limit of a sequence of continuous functions. Therefore $g(z) = f'(z)$ by Theorem 6.5.11. $\square$

Definition 11.2.9. *A complex function that can be written as a convergent power series in some open subset $U \subset \mathbb{C}$ is analytic on U .*

Remark 11.2.10. Theorem 11.2.8 says that if $f(z)$ is analytic on U , then $f(z)$ is holomorphic on U . This gives a large collection of holomorphic functions to work with. This theorem also guarantees that the derivative of any complex analytic function is also analytic, and hence holomorphic. Thus, every complex analytic function is infinitely differentiable.

Example 11.2.11.

(i) Define the function e^z by the power series

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

This series is absolutely convergent on the entire plane because for each $z \in \mathbb{C}$, the real sum $\sum_{n=0}^{\infty} \frac{|z|^n}{n!}$ converges (use the ratio test, for example). Since absolute convergence implies convergence (see Proposition 5.6.13), the series converges. This shows that e^z is holomorphic on all of $\mathbb{C}$.

(ii) Define the function $\sin(z)$ by the power series

$$\sin(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}.$$

This is absolutely convergent everywhere because the sum $\sum_{n=0}^{\infty} \frac{|z|^{2n+1}}{(2n+1)!}$ is convergent everywhere (again by the ratio test). This shows that $\sin(z)$ is holomorphic on all of $\mathbb{C}$.

(iii) Define the function $\cos(z)$ by the power series $\cos(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}$. Again, this is absolutely convergent everywhere, and hence $\cos(z)$ is holomorphic on all of $\mathbb{C}$.

The three series of the previous example are related by *Euler's formula*.

Proposition 11.2.12 (Euler's Formula). *For all $t \in \mathbb{C}$ we have*

$$e^{it} = \cos(t) + i \sin(t).$$

Proof.

$$\begin{aligned} e^{it} &= \lim_{N \rightarrow \infty} \sum_{k=0}^N \frac{i^k t^k}{k!} \\ &= \lim_{N \rightarrow \infty} \left[\sum_{k \leq N; k \text{ even}} \frac{i^k t^k}{k!} + \sum_{k \leq N; k \text{ odd}} \frac{i^k t^k}{k!} \right] \\ &= \lim_{N \rightarrow \infty} \left(\sum_{j=0}^N (-1)^j \frac{t^{2j}}{(2j)!} \right) + i \lim_{N \rightarrow \infty} \left(\sum_{\ell=0}^N (-1)^\ell \frac{t^{2\ell+1}}{(2\ell+1)!} \right) \\ &= \cos(t) + i \sin(t). \quad \square \end{aligned}$$

More discussion on the implications of Euler's formula is given in Appendix B.1.2.

Example 11.2.13. Example 11.2.11(i) shows that the function $f(z) = e^z$ is analytic, and hence holomorphic. Let's check the Cauchy-Riemann equations for f . We have $f(x, y) = e^{x+iy} = e^x \cos y + i e^x \sin y$, which gives

$$u(x, y) = e^x \cos y \quad \text{and} \quad v(x, y) = e^x \sin y.$$

Taking derivatives yields

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = e^x \cos y \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} = -e^x \sin y.$$

Thus, the Cauchy–Riemann equations hold.

Example 11.2.14. Recall that $M_n(\mathbb{C}) = \mathcal{B}(\mathbb{C}^n) = \mathcal{B}(\mathbb{C}^n; \mathbb{C}^n)$ is a Banach space (see Theorem 5.7.1). For any $A \in M_n(\mathbb{C})$ define

$$f(z) = e^{zA} = \sum_{n=0}^{\infty} \frac{(Az)^n}{n!}.$$

An argument similar to that of Example 11.2.11(i) shows that the series converges absolutely on the entire plane, and hence $f(z)$ is holomorphic on all of $\mathbb{C}$. We can use Theorem 11.2.8 to find $f'(z)$: set $z_0 = 0$, and $a_k = \frac{A^k}{k!}$. This gives

$$f'(z) = \sum_{k=1}^{\infty} k \left(\frac{A^k}{k!} \right) z^{k-1} = \sum_{k=1}^{\infty} A \left(\frac{A^{k-1}}{(k-1)!} \right) z^{k-1} = \sum_{j=0}^{\infty} A \left(\frac{A^j}{j!} \right) z^j = A e^{zA}.$$

11.3 Contour Integrals

In this section we study integration over *contours*, which are piecewise-smooth parametrized curves in $\mathbb{C}$. Although *contour integrals* are just line integrals of holomorphic functions over these piecewise-smooth curves, having a holomorphic integrand gives them useful properties that make them much simpler to evaluate than arbitrary line integrals in $\mathbb{R}^2$.

We continue to assume, throughout, that $(X, \|\cdot\|)$ is a Banach space over $\mathbb{C}$.

11.3.1 Contour Integration

Definition 11.3.1. A parametrized contour $\Gamma \subset \mathbb{C}$ is a piecewise-smooth curve in $\mathbb{C}$. That is to say, Γ consists of a sequence of C^1 -parametrized curves $\gamma_1 : [a_1, b_1] \rightarrow \mathbb{C}$, $\gamma_2 : [a_2, b_2] \rightarrow \mathbb{C}$, $\dots$, $\gamma_n : [a_n, b_n] \rightarrow \mathbb{C}$ such that the endpoint of one curve is the initial point of the next:

$$\gamma_j(b_j) = \gamma_{j+1}(a_{j+1}) \quad \text{for } j = 1, 2, \dots, n-1.$$

It is common to denote such a contour by $\Gamma = \sum_{j=1}^n \gamma_j$. We also often write $-\Gamma$ to denote the contour traversed in the opposite direction. Another common name for a contour is a path.

Definition 11.3.2. Let $U \subset \mathbb{C}$ be an open set, and let $f : U \rightarrow X$ be a continuous function. Let $\Gamma = \sum_{j=1}^n \gamma_j$ be a contour, as in the previous definition, with

$\gamma_j(t) = x_j(t) + iy_j(t)$ for each $j \in \{1, \dots, n\}$. The contour integral $\int_{\Gamma} f(z) dz$ of f over $\Gamma = \sum_{j=1}^n \gamma_j$ is the sum of the line integrals

$$\begin{aligned} \int_{\Gamma} f(z) dz &= \sum_{j=1}^n \int_{\gamma_j} f(z) dz = \sum_{j=1}^n \int_{a_j}^{b_j} f(\gamma_j(t)) \gamma_j'(t) dt \\ &= \sum_{j=1}^n \int_{a_j}^{b_j} f(\gamma_j(t)) (dx + idy) = \sum_{j=1}^n \int_{a_j}^{b_j} f(\gamma_j(t)) (x_j'(t) + iy_j'(t)) dt. \end{aligned} \quad (11.5)$$

Remark 11.3.3. A contour integral can be computed in two equivalent ways: either as $\int_{\gamma} f(z) dz$, which is a line integral in $\mathbb{C}$, or as $\int_{\gamma} f(\gamma(t)) (dx + idy)$, which is a line integral in $\mathbb{R}^2$.

Remark 11.3.4. A contour integral does not usually represent an area or a volume, but it often represents a physical quantity like work or energy. Even when the contour integral is not obviously being used to model a physical or geometric quantity, it provides an important tool for working with holomorphic functions. We use contour integrals repeatedly throughout this and the following chapters. One of the important themes of Chapter 12 is that contour integrals are very useful for understanding and manipulating linear operators.

Lemma 11.3.5. Let $\gamma : [0, 2\pi] \rightarrow \mathbb{C}$ be the circle centered at z_0 of radius r , given by $\gamma(\theta) = z_0 + re^{i\theta}$. For any $n \in \mathbb{Z}$ we have

$$\int_{\gamma} (z - z_0)^n dz = \begin{cases} 2\pi i, & n = -1, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Setting $f(z) = (z - z_0)^n$, we have

$$\begin{aligned} \int_{\gamma} f(z) dz &= \int_0^{2\pi} f(\gamma(\theta)) \gamma'(\theta) d\theta = \int_0^{2\pi} (z_0 + re^{i\theta} - z_0)^n (ire^{i\theta}) d\theta \\ &= \int_0^{2\pi} r^n e^{in\theta} (ire^{i\theta}) d\theta = ir^{n+1} \int_0^{2\pi} e^{i(n+1)\theta} d\theta. \end{aligned}$$

If $n \neq -1$, then

$$\int_{\gamma} f(z) dz = ir^{n+1} \int_0^{2\pi} e^{i(n+1)\theta} d\theta = \frac{r^{n+1}}{n+1} (e^{2\pi i(n+1)} - e^0) = 0,$$

and if $n = -1$, then $\int_{\gamma} f(z) dz = ir^0 \int_0^{2\pi} e^0 d\theta = 2\pi i$. $\square$

Remark 11.3.6. Since the contour integral is a line integral in $\mathbb{R}^2$, it is independent of parametrization, but it does depend on the orientation of the contour. If P is an oriented curve (without specific parametrization) in $\mathbb{C}$, it is common to write $\int_P f(z) dz$. Of course, to actually compute the contour integral directly from the definition requires that some parametrization of P be chosen.

If P is a closed curve in $\mathbb{C}$ with no parametrization or orientation given, it is common to assume that the orientation is counterclockwise. In this case we often write $\oint_P f(z) dz$ instead of $\int_P f(z) dz$. But if the contour is not closed, then the notions of “clockwise” and “counterclockwise” make no sense, so the orientation of P must be specified.

The next theorem is just the fundamental theorem of calculus applied to contour integrals.

Theorem 11.3.7. *If $\Gamma = \sum_{j=1}^n \gamma_j$ is a contour with endpoints z_0 and z_1 and F is a holomorphic function on an open set $U \subset \mathbb{C}$ containing Γ , then*

$$\int_{\Gamma} F'(z) dz = F(z_1) - F(z_0).$$

Proof. For each smooth curve $\gamma_j : [a_j, b_j] \rightarrow \mathbb{C}$ we have

$$\int_{\gamma_j} F'(z) dz = \int_{a_j}^{b_j} F'(\gamma_j(t)) \gamma_j'(t) dt = \int_{a_j}^{b_j} \frac{d}{dt} F(\gamma_j(t)) dt = F(\gamma_j(b_j)) - F(\gamma_j(a_j)).$$

Since $\gamma_j(b_j) = \gamma_{j+1}(a_{j+1})$, the sum (11.5) collapses to give $F(\gamma_n(b_n)) - F(\gamma_1(a_1))$. Since $z_0 = \gamma_1(a_1)$ and $z_1 = \gamma_n(b_n)$, we have the desired result. $\square$

Remark 11.3.8. The theorem above implies that the value of the contour integral $\int_{\Gamma} F'(z) dz$ depends only on the endpoints of Γ and is independent of the path itself, provided Γ lies entirely in U .

Corollary 11.3.9. *If Γ is a closed contour in $\mathbb{C}$ and F is a holomorphic function on an open set $U \subset \mathbb{C}$ containing Γ , then*

$$\int_{\Gamma} F'(z) dz = 0.$$

Example 11.3.10. In the case of the integral $\int_{\gamma} (z - z_0)^n dz$ in Lemma 11.3.5, if $n \neq -1$, then we have $(z - z_0)^n = \frac{d}{dz} \frac{(z - z_0)^{n+1}}{n+1}$, so the corollary implies $\int_{\gamma} (z - z_0)^n dz = 0$.

One might guess that the function $(z - z_0)^{-1}$ should be the derivative of $\log(z - z_0)$, but that is not correct because $\log(z - z_0)$ is not a function on $\mathbb{C}$. To see this, recall that for $x \in (0, \infty) \subset \mathbb{R}$ the function $\log(x)$ is the inverse function of e^y ; that is, $y = \log(x)$ means that y is the unique number such that $e^y = x$. This is not a problem in $(0, \infty)$ because e^y is injective as a map from $\mathbb{R}$ to $(0, \infty)$, but e^z is not injective as a map from $\mathbb{C}$ to $\mathbb{C}$; for example, $e^{2\pi i k} = e^0 = 1$ for all $k \in \mathbb{Z}$. Since e^z is not injective, it can have no inverse.

We can get around the failure to be injective by restricting the domain (just as with the inverse sine and inverse cosine); but even so, we cannot define

$\log(z)$ as a single holomorphic function on a closed contour that encircles the origin, and $(z - z_0)^{-1}$ is not a derivative of any single function on such a contour. This is the reason that the corollary does not force the integral $\int_{\gamma} (z - z_0)^{-1} dz$ to vanish.

11.3.2 The Cauchy–Goursat Theorem

In this section we consider contour integrals where the contour Γ forms a simple closed curve in a simply connected set U . When we integrate a function f that is holomorphic on U , the value of the contour integral is always zero, regardless of the contour. This theorem, called the Cauchy–Goursat theorem, is one of the fundamental results in complex analysis.

Theorem 11.3.11 (Cauchy–Goursat Theorem). *If $f: U \rightarrow X$ is holomorphic on a simply connected set $U \subset \mathbb{C}$, then*

$$\int_{\Gamma} f(z) dz = 0 \quad (11.6)$$

for any simple closed contour Γ in U .

Remark 11.3.12. Compare the Cauchy–Goursat theorem to Corollary 11.3.9. This theorem says that if we add the hypothesis that U is simply connected, then we no longer need the requirement that the integrand be a derivative of a holomorphic function—it suffices for the integrand to be holomorphic on U .

Remark 11.3.13. Let $f: U \rightarrow X$. If we can write $f = u + iv$, where u and v are C^1 functions from $U \subset \mathbb{R}^2$ to X (taken as a Banach space over $\mathbb{R}$), then since U is simply connected, we could use Green’s theorem as follows. If R is the region that has Γ as its boundary, then

$$\begin{aligned} \int_{\Gamma} f(z) dz &= \int_{\Gamma} [u(x, y) + iv(x, y)] (dx + idy) \\ &= \int_{\Gamma} (u + iv) dx + (-v + iu) dy \\ &= \iint_R \left[\frac{\partial}{\partial x}(-v + iu) - \frac{\partial}{\partial y}(u + iv) \right] dx dy \\ &= \iint_R \left[-\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) + i \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) \right] dx dy. \end{aligned}$$

The last integral vanishes by the Cauchy–Riemann equations (Theorem 11.1.4).

Unfortunately, we cannot use Green’s theorem here because in general we do not know that f can be written as $u + iv$, where u and v are C^1 . We show in Corollary 11.4.8 that u and v must, in fact, be C^∞ whenever $f = u + iv$ is holomorphic, but that proof depends on the Cauchy–Goursat theorem, so we cannot use it here.

We give a complete proof of the Cauchy–Goursat theorem in the next section.

Example 11.3.14. Lemma 11.3.5 showed that when Γ is a circle around z_0 and $n \in \mathbb{N}$, then $\int_{\Gamma} (z - z_0)^n dz = 0$. Since $(z - z_0)^n$ is holomorphic on the entire plane when $n \in \mathbb{N}$, this is consistent with the Cauchy–Goursat theorem.

In the case that $n < 0$ the Cauchy–Goursat theorem does not immediately apply, because the function $(z - z_0)^n$ is not continuous and hence not holomorphic at $z = z_0$, and the punctured plane $\mathbb{C} \setminus \{z_0\}$ is not simply connected.

Remark 11.3.15. The Cauchy–Goursat theorem implies that contour integrals of any holomorphic function are path independent on a simply connected domain. That is to say, if γ_1 and γ_2 are two contours in a simply connected domain U with the same starting and ending points, then the contour $\gamma_1 - \gamma_2$ is a closed contour in U , so

$$0 = \int_{\gamma_1 - \gamma_2} f(z) dz = \int_{\gamma_1} f(z) dz - \int_{\gamma_2} f(z) dz.$$

Hence

$$\int_{\gamma_1} f(z) dz = \int_{\gamma_2} f(z) dz.$$

Example 11.3.16. Lemma 11.3.5 shows that $\int_{\gamma_0} (z - z_0)^{-1} dz = 2\pi i$, where γ_0 is a circle around z_0 , traversed once in a counterclockwise direction. Consider now a different path γ_1 that also wraps once around the point z_0 , as in Figure 11.1. Adding a segment σ (a “cut”) from the starting (and ending) point of γ_1 to the starting (and ending) point of γ_0 , the contour $\Gamma = \gamma_1 + \sigma - \gamma_0 - \sigma$ encloses a simply connected region; hence the integral $\int_{\Gamma} (z - z_0)^{-1} dz$ vanishes, and we have $\int_{\gamma_0} (z - z_0)^{-1} dz = \int_{\gamma_1} (z - z_0)^{-1} dz$, since the contribution from σ cancels.

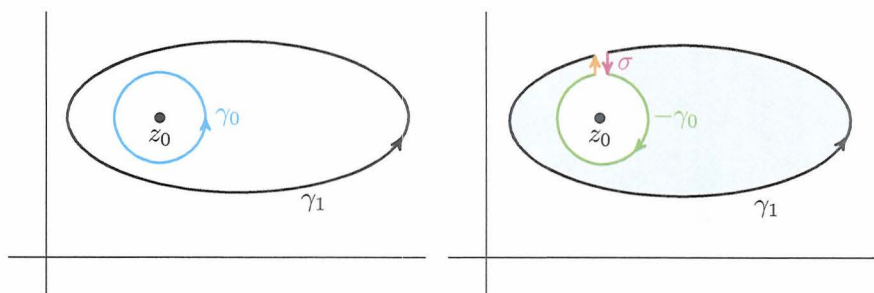


Figure 11.1. If f is a function that is holomorphic everywhere but at z_0 , the integrals of f over γ_1 and γ_0 are same. To see this, add a little cut σ (red) so that the region (shaded blue) bounded by $\gamma_1 + \sigma - \gamma_0 - \sigma$ is simply connected. The integral over $\gamma_1 + \sigma - \gamma_0 - \sigma$ is 0 by the Cauchy–Goursat theorem (Theorem 11.3.11), and the contributions from σ and $-\sigma$ cancel.

11.3.3 * Proof of the Cauchy–Goursat Theorem

Remark 11.3.15 shows that when $f = u + iv$ is holomorphic, if u and v are C^1 , then the Cauchy–Goursat theorem follows from Green’s theorem. Goursat was the first to prove this theorem without assuming that u and v are C^1 . In this section we give a proof of this theorem.

Lemma 11.3.17 (Cauchy–Goursat for “Good” Contours). *Let $U \subset \mathbb{C}$ be a simply connected set. As in Definition 9.5.1, for each $k \in \mathbb{N}$ let Q_k be the set of points in $\mathbb{C}$ with (real and imaginary) coordinates of the form $c/2^k$ for $c \in \mathbb{Z}$, and let Θ_k be the set of squares of side-length 2^{-k} in $\mathbb{C}$ with coordinates in Q_k . Let Γ be a closed contour in U with interior Ω . Let $\overline{\Omega} = \Omega \cup \Gamma$.*

Let $f: U \rightarrow X$ be holomorphic on U . Assume there exists a $K > 0$ such that for every $k \geq K$ and for every $R \in \Theta_k$, the intersection $\overline{\Omega} \cap R$ has a finite number of boundary components.

In this case we have

$$\int_{\Gamma} f(z) dz = 0. \quad (11.7)$$

Note that the conditions of the lemma hold for every closed contour consisting of a finite sum $\Gamma = \gamma_1 + \cdots + \gamma_n$, where each γ_i is linear. We call such paths *polygonal*.

Proof. Note that the interior Ω of Γ is bounded, so $\overline{\Omega} = \Omega \cup \Gamma$ is compact. Because f is holomorphic on U , for each $\varepsilon > 0$ and for each $z \in \overline{\Omega}$ there exists a $\delta_z > 0$ such that $\|f(w) - f(z) - f'(z)(w - z)\|_X \leq \varepsilon|z - w|$ whenever $|z - w| < \delta_z$. Moreover, we may assume that $B(z, \delta_z) \subset U$. The collection of all balls $B(z, \delta_z)$ covers $\overline{\Omega}$, so the cover has a Lebesgue number, which we denote by δ (see Theorem 5.5.11). Choose $m \in \mathbb{N}$ such that $\sqrt{2}/2^m < \delta$. For every point z of $\overline{\Omega}$ and for every $R \in \Theta_m$ containing z , every $w \in R$ satisfies $|w - z| \leq \sqrt{2}/2^m < \delta$, so we have

$$\|f(w) - f(z) - f'(z)(w - z)\| \leq \varepsilon|z - w|. \quad (11.8)$$

Since $\overline{\Omega}$ is compact, it is covered by a finite number of squares $R_1, \dots, R_n \in \Theta_m$. Discard from this list any square that does not meet $\overline{\Omega}$. For each $i \in \{1, \dots, n\}$ choose a $z_i \in R_i \cap \overline{\Omega}$ and define a new function on R_i by

$$g_i(w) = \begin{cases} 0 & \text{if } w = z_i, \\ \frac{f(w) - f(z_i)}{w - z_i} - f'(z_i) & \text{if } w \neq z_i. \end{cases}$$

By (11.8) we have $\|g_i(w)\| < \varepsilon$, provided $w \in R_i$.

For each $i \in \{1, \dots, n\}$ let C_i denote the positively oriented boundary of $R_i \cap \overline{\Omega}$. By hypothesis, the contour C_i must consist of only finitely many connected components, and each of them is a closed contour. We have

$$\sum_{i=1}^n \int_{C_i} f(z) dz = \int_{\Gamma} f(z) dz,$$

since any oriented segment of a side of R_i that is not part of Γ is also contained in exactly one other R_j , but with the opposite orientation in R_j , and thus they cancel out in the sum.

By definition of g_i we have

$$\int_{C_i} f(w) dw = \int_{C_i} f(z_i) - z_i f'(z_i) + w f'(z_i) + (w - z_i) g_i(w) dw.$$

But by Corollary 11.3.9 we have $\int_{C_i} dw = 0$ and $\int_{C_i} w dw = 0$, and thus

$$\int_{C_i} f(w) dw = \int_{C_i} (w - z_i) g_i(w) dw.$$

Summing over all $i \in \{1, \dots, n\}$ and using the fact that both $|w - z_i| < \sqrt{2}/2^m$ and $\|g_i(w)\| < \varepsilon$ for all $w \in R_i$, we get

$$\left\| \int_{\Gamma} f(w) dw \right\| = \left\| \sum_{i=1}^n \int_{C_i} (w - z_i) g_i(w) dw \right\| < \sum_{i=1}^n \varepsilon 2^{-m} \sqrt{2} (\text{len}(C_i)).$$

The sum of the lengths of the contours C_i cannot be longer than the length of Γ plus the sum of the lengths of all the sides of the R_i (that is, $4 \cdot 2^{-m}$), so we have

$$\left\| \int_{\Gamma} f(w) dw \right\| < \varepsilon 2^{-m} \sqrt{2} (4n2^{-m} + \text{len}(\Gamma)) = \varepsilon \sqrt{2} (4A + 2^{-m} \text{len}(\Gamma)), \quad (11.9)$$

where A is the total area of all the squares $R_1 \cup \dots \cup R_n$. Since $\overline{\Omega}$ is compact and measurable, the area A approaches the (finite) measure of $\overline{\Omega}$ as $m \rightarrow \infty$. Since ε is arbitrary, we may make the right side of (11.9) as small as we please, and hence the left side must vanish. $\square$

To extend the result to a general closed contour we first prove one more lemma.

Lemma 11.3.18. *Let $\gamma : [a, b] \rightarrow U$ be C^1 , and let f be holomorphic on U . For any $\varepsilon > 0$ there is a polygonal path $\sigma = \sum_{k=1}^n \sigma_k \subset U$ with $\sigma(a) = \gamma(a)$ and $\sigma(b) = \gamma(b)$ such that*

$$\left\| \int_{\sigma} f(z) dz - \int_{\gamma} f(z) dz \right\| < \varepsilon.$$

Proof. Since γ is compact, the distance $\rho = d(\gamma, U^c)$ to U^c must be positive (see Exercise 5.33). Let Z be the compact set $Z = \{z \in U \mid d(z, \gamma) \leq \rho/2\} \subset U$. Since f is holomorphic, it is uniformly continuous on Z .

Since γ is C^1 it has a well-defined arclength $L = \int_{\gamma} |dz|$. For every $\varepsilon > 0$, choose $\delta > 0$ so that for all $z, w \in Z$ we have

$$\|f(z) - f(w)\| < \min(\varepsilon/3L, \rho/2),$$

provided $|z - w| < \delta$. The path γ is uniformly continuous on $[a, b]$, so there exists an $\eta > 0$ such that

$$|\gamma(t) - \gamma(s)| < \delta$$

for all $s, t \in [a, b]$, provided $|s - t| < \eta$. Finally, γ' is uniformly continuous on $[a, b]$, so letting $F = \sup_{t \in [a, b]} \|f(\gamma(t))\|$, there exists $\xi > 0$ so that

$$|\gamma'(t) - \gamma'(s)| < \frac{\varepsilon}{3(b-a)F}, \quad (11.10)$$

whenever $|s - t| < \xi$. Choose a partition $a = t_0 < t_1 < \cdots < t_n = b$ of $[a, b]$ such that $|t_k - t_{k-1}| < \min(\eta, \xi)$ for every k .

Let

$$S = \sum_{k=1}^n f(\gamma(t_k))[\gamma(t_k) - \gamma(t_{k-1})],$$

and let $\sigma_k(t)$ with $t \in [t_{k-1}, t_k]$ be the line segment from $\gamma(t_{k-1})$ to $\gamma(t_k)$; that is,

$$\sigma_k(t) = \frac{1}{t_k - t_{k-1}} [t(\gamma(t_k) - \gamma(t_{k-1})) + t_k\gamma(t_{k-1}) - t_{k-1}\gamma(t_k)].$$

Taking $\sigma = \sum_{k=1}^n \sigma_k(t)$ gives a polygonal path lying entirely in Z with endpoints $\gamma(a)$ and $\gamma(b)$.

We have

$$\begin{aligned} \left\| \int_{\sigma} f(z) dz - S \right\| &= \left\| \int_{\sigma} f(z) dz - \sum_{k=1}^n f(\gamma(t_k))[\gamma(t_k) - \gamma(t_{k-1})] \right\| \\ &= \left\| \int_{\sigma} f(z) dz - \sum_{k=1}^n \int_{\gamma(t_{k-1})}^{\gamma(t_k)} f(\gamma(t_k)) dz \right\| \\ &= \left\| \sum_{k=1}^n \int_{\gamma(t_{k-1})}^{\gamma(t_k)} [f(z) - f(\gamma(t_k))] dz \right\| \\ &\leq \sum_{k=1}^n \int_{\gamma(t_{k-1})}^{\gamma(t_k)} \|f(z) - f(\gamma(t_k))\| |dz| \\ &< \frac{\varepsilon}{3L} \sum_{k=1}^n \int_{\gamma(t_{k-1})}^{\gamma(t_k)} |dz| = \frac{\varepsilon}{3}. \end{aligned} \quad (11.11)$$

By the mean value theorem (Theorem 6.5.1), for each k there exists a $t_k^* \in [t_{k-1}, t_k]$ such that

$$S = \sum_{k=1}^n f(\gamma(t_k))[\gamma(t_k) - \gamma(t_{k-1})] = \sum_{k=1}^n f(\gamma(t_k))\gamma'(t_k^*)[t_k - t_{k-1}]. \quad (11.12)$$

Since $f \circ \gamma$ is continuous, it is Riemann integrable, so after further refining the partition, if necessary, we may assume that

$$\left\| \int_{\gamma} f(z) dz - \sum_{k=1}^n f(\gamma(t_k))\gamma'(t_k)[t_k - t_{k-1}] \right\| < \frac{\varepsilon}{3}. \quad (11.13)$$

Combining (11.10), (11.11), (11.12), and (11.13) with the triangle inequality gives

$$\begin{aligned} \left\| \int_{\gamma} f(z) dz - \int_{\sigma} f(z) dz \right\| &\leq \left\| \int_{\gamma} f(z) dz - \sum_{k=1}^n f(\gamma(t_k))\gamma'(t_k)(t_k - t_{k-1}) \right\| \\ &\quad + \left\| \sum_{k=1}^n f(\gamma(t_k))\gamma'(t_k)(t_k - t_{k-1}) - S \right\| + \left\| S - \int_{\sigma} f(z) dz \right\| \\ &< \frac{\varepsilon}{3} + \left\| \sum_{k=1}^n f(\gamma(t_k))[\gamma'(t_k) - \gamma'(t_k^*)](t_k - t_{k-1}) \right\| + \frac{\varepsilon}{3} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2\varepsilon}{3} + F \sum_{k=1}^n |\gamma'(t_k) - \gamma'(t_k^*)|(t_k - t_{k-1}) \\
&< \frac{2\varepsilon}{3} + \frac{\varepsilon}{3(b-a)} \sum_{k=1}^n (t_k - t_{k-1}) = \varepsilon. \quad \square
\end{aligned}$$

The proof of Cauchy–Goursat is now straightforward.

Proof of Theorem 11.3.11. Let $\Gamma \subset U$ be a closed contour, with $\Gamma = \sum_{i=1}^m \gamma_i$, where $\gamma_1 : [a_1, b_1] \rightarrow \mathbb{C}$, $\gamma_2 : [a_2, b_2] \rightarrow \mathbb{C}$, $\dots$, $\gamma_m : [a_m, b_m] \rightarrow \mathbb{C}$ are all C^1 . For any $\varepsilon > 0$, Lemma 11.3.18 guarantees there are polygonal paths $\sigma_1, \dots, \sigma_m$ such that $\sigma_i(a_i) = \gamma_i(a_i)$ and $\sigma_i(b_i) = \gamma_i(b_i)$ for every i and such that $\|\int_{\gamma_i} f(z) dz - \int_{\sigma_i} f(z) dz\| < \varepsilon/m$. Letting σ be the polygonal path $\sigma = \sum_{i=1}^m \sigma_i$ gives

$$\left\| \int_{\gamma} f(z) dz - \int_{\sigma} f(z) dz \right\| < \varepsilon.$$

But σ is a closed polygonal contour, and therefore Lemma 11.3.17 applies to give

$$\int_{\sigma} f(z) dz = 0;$$

hence,

$$\left\| \int_{\gamma} f(z) dz \right\| < \varepsilon.$$

Since this holds for every $\varepsilon > 0$ we must have

$$\left\| \int_{\gamma} f(z) dz \right\| = 0. \quad \square$$

11.4 Cauchy's Integral Formula

Cauchy's integral formula is probably the single most important result in all of complex analysis. Among other things, it implies that holomorphic functions are infinitely differentiable (smooth)—something that is certainly not true for arbitrary differentiable functions from $\mathbb{R}^2$ to $\mathbb{R}^2$. It also is used to show that every holomorphic function can be represented as a convergent power series in a small neighborhood. That is to say, all holomorphic functions are analytic (see Definition 11.2.9). This is a very powerful result that vastly simplifies many problems.

Cauchy's integral formula is also useful for proving several other important results about complex functions throughout this chapter and for proving important results about linear operators throughout the next chapter.

We continue to assume that $(X, \|\cdot\|)$ is a Banach space over $\mathbb{C}$.

11.4.1 Cauchy's Integral Formula

Theorem 11.4.1 (Cauchy's Integral Formula). *Let X be a Banach space over $\mathbb{C}$, and let $f : U \rightarrow X$ be holomorphic on an open, simply connected domain $U \subset \mathbb{C}$.*

Let γ be a simple closed contour lying entirely in U , traversed once counterclockwise. For any z_0 in the interior of γ , we have

$$f(z_0) = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z - z_0} dz. \quad (11.14)$$

Proof. If σ is any circle centered at z_0 and lying entirely in the interior of γ , then using the Cauchy–Goursat theorem with the same argument as in Example 11.3.16 and Figure 11.1, we have that

$$\oint_{\sigma} \frac{f(z)}{z - z_0} dz = \oint_{\gamma} \frac{f(z)}{z - z_0} dz.$$

Therefore, it suffices to prove the result in the case that γ is some circle of sufficiently small radius r .

By Lemma 11.3.5 we have

$$f(z_0) = f(z_0) \frac{1}{2\pi i} \oint_{\gamma} \frac{1}{z - z_0} dz = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z_0)}{z - z_0} dz,$$

which gives

$$\begin{aligned} \left\| \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z - z_0} dz - f(z_0) \right\| &= \left\| \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z) - f(z_0)}{z - z_0} dz \right\| \\ &= \left\| \frac{1}{2\pi i} \int_0^{2\pi} \frac{f(z_0 + re^{it}) - f(z_0)}{re^{it}} ire^{it} dt \right\| \\ &= \left\| \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{it}) - f(z_0) dt \right\| \\ &\leq \sup_{|z - z_0| = r} \|f(z) - f(z_0)\|. \end{aligned}$$

Since f is holomorphic in U , it is continuous at z_0 ; so, for any $\varepsilon > 0$ there is a choice of $\delta > 0$ such that $\|f(z) - f(z_0)\| < \varepsilon$ whenever $|z - z_0| \leq \delta$. Therefore, choosing $0 < r < \delta$, we have

$$\left\| \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z - z_0} dz - f(z_0) \right\| < \varepsilon.$$

Since this holds for every positive ε , the result follows. $\square$

Example 11.4.2.

- (i) Let Γ be any simple closed contour around 0. Since e^z is holomorphic everywhere, we compute

$$\oint_{\Gamma} \frac{e^z}{z} dz = 2\pi i e^0 = 2\pi i.$$

(ii) Consider the integral

$$\oint_{\Gamma} \frac{\cos(z)}{z + z^4} dz,$$

where Γ is a circle of radius strictly less than one around the origin. The function $f(z) = \cos(z)/(1+z^3)$ is holomorphic away from the three roots of $1+z^3=0$, which all have norm 1. So inside of the open ball $B(0,1)$ the function $f(z)$ is holomorphic, and Γ lies inside of $B(0,1)$. We now compute

$$\oint_{\Gamma} \frac{\cos(z)}{z + z^4} dz = \oint_{\Gamma} \frac{f(z)}{z} dz = 2\pi i f(0) = 2\pi i.$$

Corollary 11.4.3 (Gauss's Mean Value Theorem). *Let f be holomorphic in a simply connected domain U , such that U contains the circle $C = \{z \in \mathbb{C} : |z - z_0| = r\}$ of radius r with center z_0 . We have*

$$f(z_0) = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{it}) dt.$$

Proof. Parametrize C in the usual way, $\gamma(t) = z_0 + re^{it}$ for $t \in [0, 2\pi]$, and apply the Cauchy integral formula:

$$\begin{aligned} f(z_0) &= \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - z_0} dz \\ &= \frac{1}{2\pi i} \int_0^{2\pi} \frac{f(z_0 + re^{it})}{re^{it}} ire^{it} dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{it}) dt. \quad \square \end{aligned}$$

Remark 11.4.4. Gauss's mean value theorem says that the average (mean) value of a holomorphic function on a circle is the value of the function at the center of the circle. This is much more precise than the other mean value theorems we have encountered, which just say that the average value over a set is achieved somewhere in the set, but do not specify where that mean value is attained.

11.4.2 Riemann's Theorem and Cauchy's Differentiation Formula

Cauchy's integral formula can be generalized to derivatives as well. As a first step, we prove Riemann's theorem.

Theorem 11.4.5 (Riemann's Theorem). *Let f be continuous on a closed contour γ . For every positive integer n the function*

$$F_n(w) = \oint_{\gamma} \frac{f(z)}{(z - w)^n} dz$$

is holomorphic at all w in the complement of γ (both the interior and exterior components), and its derivative satisfies

$$F'_n(w) = nF_{n+1}(w).$$

Proof. The result follows immediately from Leibniz's integral rule (Theorem 8.6.9), which allows us to pass the derivative through the integral. For every n we have

$$\begin{aligned}\frac{d}{dw} F_n(w) &= \frac{d}{dw} \oint_{\gamma} \frac{f(z)}{(z-w)^n} dz \\ &= \oint_{\gamma} \frac{\partial}{\partial w} \frac{f(z)}{(z-w)^n} dz \\ &= \oint_{\gamma} \frac{nf(z)}{(z-w)^{n+1}} dz \\ &= nF_{n+1}(w). \quad \square\end{aligned}$$

Corollary 11.4.6 (Cauchy's Differentiation Formula). *Let f be holomorphic on an open, simply connected domain $U \subset \mathbb{C}$. Let γ be a simple closed contour lying entirely in U , traversed once counterclockwise. For any w in the interior of γ , we have*

$$f^{(k)}(w) = \frac{k!}{2\pi i} \oint_{\gamma} \frac{f(z)}{(z-w)^{k+1}} dz. \quad (11.15)$$

Proof. Proceed by induction on k . The case of $k = 0$ is Cauchy's integral formula (Theorem 11.4.1). Assume that

$$f^{(k-1)}(w) = \frac{(k-1)!}{2\pi i} \int_{\gamma} \frac{f(z)}{(z-w)^k} dz.$$

Riemann's theorem gives

$$\frac{d}{dw} f^{(k-1)}(w) = \frac{k!}{2\pi i} \int_{\gamma} \frac{f(z)}{(z-w)^{k+1}} dz,$$

which completes the induction step. $\square$

Nota Bene 11.4.7. The following, equivalent form of (11.15) is often useful

$$\int_{\gamma} \frac{f(z)}{(z-w)^k} dz = \frac{2\pi i f^{(k-1)}(w)}{(k-1)!} \quad (\text{for } w \text{ inside } \Gamma). \quad (11.16)$$

Corollary 11.4.8. *If f is a function that is holomorphic on an open set U , then the derivative of f is holomorphic on U , and f is infinitely differentiable on U .*

Remark 11.4.9. This last corollary is yet another demonstration of the fact that holomorphic functions are very special, because there are many functions from $\mathbb{R}^2$ to $\mathbb{R}^2$ that are differentiable but not twice differentiable, whereas any function from $\mathbb{C}$ to $\mathbb{C}$ that is differentiable as a complex function (holomorphic) is also infinitely differentiable.

Example 11.4.10.

- (i) To compute the contour integral

$$\oint_C \frac{\cos(z)}{2z^3} dz,$$

where C is any simple closed contour enclosing 0, use the Cauchy differentiation formula with $f(z) = \cos(z)/2$ to get

$$\oint_C \frac{\cos(z)}{2z^3} dz = \frac{2\pi i}{2!} f''(0) = -\frac{\pi i}{2}.$$

- (ii) Consider the integral

$$\oint_{\Gamma} \frac{\sin(z)}{z^3 + z^5} dz,$$

where Γ is a circle of radius strictly less than one around the origin. Note that $f(z) = \frac{\sin(z)}{(1+z^2)}$ is a holomorphic function inside the circle Γ because it can be written as a product of two holomorphic functions $\sin(z)$ and $1/(1+z^2)$. Therefore, we have

$$\begin{aligned} \oint_{\Gamma} \frac{\sin(z)}{z^3 + z^5} dz &= \oint_{\Gamma} \frac{f(z)}{z^3} dz \\ &= \frac{2\pi i}{2!} f''(0) = \pi i \left. \frac{d^2}{dz^2} \frac{\sin(z)}{1+z^2} \right|_{z=0} = 0. \end{aligned}$$

- (iii) Consider the integral

$$\oint_{\gamma} \frac{1}{(z^2 - 1)^2(z - 1)} dz = \oint_{\gamma} \frac{1}{(z - 1)^3(z + 1)^2} dz,$$

where γ is a circle centered at 0 with radius strictly greater than 1. Since the integrand fails to be holomorphic at both -1 and $+1$, and since γ encloses both 1 and -1 , Cauchy's formulas do not apply with the integral as given.

To fix this problem, let γ_1 and γ_2 be circles of radius less than 1 with centers equal to -1 and 1 , respectively. Applying the cutting trick used in Figure 11.1 to the contours $\gamma_1 + \gamma_2$ versus γ (see Figure 11.2) gives

$$\begin{aligned} \oint_{\gamma} \frac{1}{(z - 1)^3(z + 1)^2} dz &= \oint_{\gamma_1 + \gamma_2} \frac{1}{(z - 1)^3(z + 1)^2} dz \\ &= \oint_{\gamma_1} \frac{1}{(z - 1)^3(z + 1)^2} dz + \oint_{\gamma_2} \frac{1}{(z - 1)^3(z + 1)^2} dz. \end{aligned}$$

But inside γ_1 the function $f_1 = \frac{1}{(z-1)^3}$ is holomorphic, so

$$\oint_{\gamma_1} \frac{1}{(z^2-1)^2(z-1)} dz = \oint_{\gamma_1} \frac{f_1(z)}{(z+1)^2} dz = 2\pi i f_1'(-1) = -\frac{3\pi i}{8}.$$

Similarly, inside γ_2 the function $f_2 = \frac{1}{(z+1)^2}$ is holomorphic, so

$$\oint_{\gamma_2} \frac{1}{(z^2-1)^2(z-1)} dz = \oint_{\gamma_2} \frac{f_2(z)}{(z-1)^3} dz = \pi i f_2''(1) = \frac{3\pi i}{8}.$$

Therefore the integral is equal to 0.

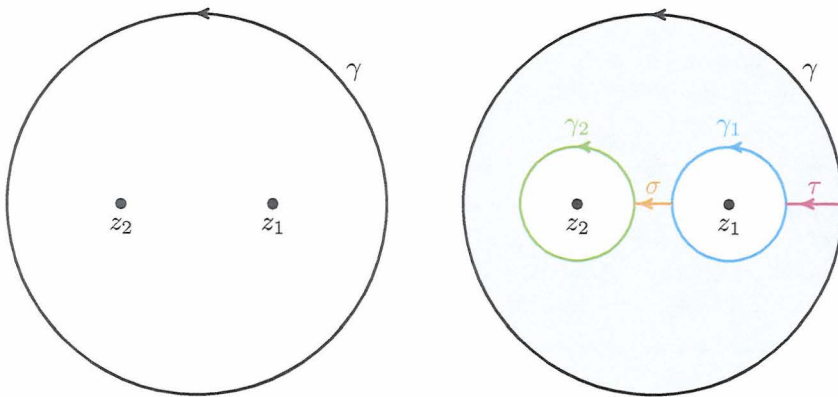


Figure 11.2. When integrating a function f that is holomorphic everywhere inside of γ except at z_1 and z_2 , we can replace the contour γ with two smaller circles: γ_1 around z_1 and γ_2 around z_2 . To do this, add a short horizontal cut τ (red) from γ to γ_1 and another σ (orange) from γ to γ_2 . Removing σ , τ , and the interiors of γ_1 and γ_2 from the interior of γ leaves a simply connected region (shaded blue), so Cauchy's integral formula guarantees that the integral of f over γ is the same as the sum of integrals over τ , the upper half of γ_1 , σ , γ_2 , $-\sigma$, the lower half of γ_1 , and $-\tau$. But the integrals involving τ and σ cancel, and the two halves of γ_1 combine to give $\oint_{\gamma} f dz = \oint_{\gamma_1} f dz + \oint_{\gamma_2} f dz$. See Example 11.4.10(iii) for a specific case of this.

11.5 Consequences of Cauchy's Integral Formula

Cauchy's integral formula has many important consequences, including the fact that every bounded holomorphic function is constant (Liouville's theorem), the fact that holomorphic functions on a disk attain their maximum norm on the boundary of the disk (the maximum modulus principle), and the fact that every polynomial with coefficients in $\mathbb{C}$ has a root in $\mathbb{C}$ (the fundamental theorem of algebra). We discuss all three of these in this section.

11.5.1 Liouville's Theorem

Theorem 11.5.1 (Liouville's Theorem). *If $f: \mathbb{C} \rightarrow X$ is holomorphic and bounded on all of $\mathbb{C}$, then f is constant.*

Proof. Let $M \in \mathbb{R}$ be such that $\|f(z)\| < M$ for all $z \in \mathbb{C}$. Choose any $z_0 \in \mathbb{C}$ and let γ be a circle of radius R centered at z_0 . Then

$$\begin{aligned} \|f'(z_0)\| &= \frac{1}{2\pi} \left\| \int_{\gamma} \frac{f(z)}{(z - z_0)^2} dz \right\| \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} \frac{\|f(z_0 + Re^{it})\|}{|Re^{it}|^2} |Re^{it}| dt \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} \frac{M}{R} dt = \frac{M}{R}. \end{aligned}$$

This holds for every $R > 0$, so as $R \rightarrow \infty$, we see that $\|f'(z_0)\| = 0$. Thus f is constant by Proposition 11.2.4. $\square$

Example 11.5.2. The functions $\sin(z)$ and $\cos(z)$ are bounded if $z \in \mathbb{R}$, but since they are not constant, Liouville's theorem guarantees they cannot be bounded on all of $\mathbb{C}$. That is, for every $M > 0$, there must be a $z \in \mathbb{C}$ such that $|\sin(z)| > M$, and similarly for $\cos(z)$.

Corollary 11.5.3. *If $f: \mathbb{C} \rightarrow \mathbb{C}$ is holomorphic on all of $\mathbb{C}$, and if $|f|$ is uniformly bounded away from zero (meaning there is an $\varepsilon > 0$ such that $|f(z)| > \varepsilon$ for all $z \in \mathbb{C}$), then f is constant.*

Proof. If $|f| > \varepsilon$, then $|1/f| < 1/\varepsilon$. Moreover, since f is entire and nonzero, $1/f$ is also entire. Hence, by Liouville's theorem, $1/f$ is constant, which implies f is constant. $\square$

11.5.2 The Fundamental Theorem of Algebra

The fundamental theorem of algebra is a very significant consequence of Liouville's theorem.

Theorem 11.5.4 (Fundamental Theorem of Algebra). *Every nonconstant polynomial over $\mathbb{C}$ has at least one root in $\mathbb{C}$.*

Proof. Let f be any nonconstant polynomial of degree $k > 0$, let c be the coefficient of the degree- k term, and let $p = f/c$, so

$$p(z) = z^k + a_{k-1}z^{k-1} + \cdots + a_1z + a_0.$$

It suffices to show that $p(z)$ has a root. Let $a = \max\{|a_{k-1}|, |a_{k-2}|, \dots, |a_0|\}$. If $a = 0$, then $p(z) = z^k$ and $z = 0$ is a root of p ; thus, we may assume $a > 0$. Suppose that $p(z)$ has no roots, and let $R = \max\{(k+1)a, 1\}$. If $|z| \geq R$, then

$$\begin{aligned} |p(z)| &\geq |z|^k - (|a_{k-1}||z|^{k-1} + \dots + |a_1||z| + |a_0|) \\ &\geq |z|^k - ka|z|^{k-1} \\ &= |z|^{k-1}(|z| - ka) \\ &\geq |z| - ka \geq R - ka \geq a. \end{aligned}$$

Thus, $p(z)$ is uniformly bounded away from zero on the exterior of $\overline{B(0, R)}$. Since no roots exist for $|z| \leq R$, compactness and continuity imply that $p(z)$ is also uniformly bounded away from zero on the interior of $\overline{B(0, R)}$. By Corollary 11.5.3, we must have that $p(z)$ is constant, which is a contradiction. $\square$

11.5.3 The Maximum Modulus Principle

Recall that the extreme value theorem (Corollary 5.5.7) guarantees every continuous, real-valued function on a compact set attains its maximum. The maximum modulus principle is a sort of converse for holomorphic functions. It says that nonconstant holomorphic functions never achieve their maximum norm on open sets. That is, there are no holomorphic functions whose norms have a local maximum. This may be surprising, since there are lots of nonconstant real functions that have many local maxima.

Theorem 11.5.5 (Maximum Modulus Principle). *Let $f: U \rightarrow \mathbb{C}$ be $\mathbb{C}$ -valued, holomorphic, and not constant on an open, path-connected set U . The modulus $|f(z)|$ of f never attains its supremum on U .*

The proof depends on the following lemma.

Lemma 11.5.6. *Let f be holomorphic on an open, path-connected set U . If $|f|$ attains its maximum at $z_0 \in U$, then $|f(z)|$ is constant in every open ball $B(z_0, r)$ whose closure $\overline{B(z_0, r)}$ is entirely contained in U .*

Proof. For any $r > 0$ with $\overline{B(z_0, r)} \subset U$, Gauss's mean value theorem (Corollary 11.4.3) implies

$$\begin{aligned} |f(z_0)| &= \frac{1}{2\pi} \left| \int_0^{2\pi} f(z_0 + re^{it}) dt \right| \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + re^{it})| dt. \end{aligned}$$

But since $|f|$ attains its maximum at z_0 we have

$$|f(z_0)| \geq |f(z_0 + re^{it})| \tag{11.17}$$

for every $t \in [0, 2\pi]$, so

$$|f(z_0)| \geq \frac{1}{2\pi} \int_0^{2\pi} |f(z_0 + re^{it})| dt,$$

and hence equality holds. And thus we see

$$0 = \frac{1}{2\pi} \int_0^{2\pi} |f(z_0)| - |f(z_0 + re^{it})| dt.$$

But by (11.17) the integrand is nonnegative; hence it must be zero. Therefore, $|f(z)| = |f(z_0)|$ whenever $|z - z_0| = r$. For every $\varepsilon \leq r$ we also have $\overline{B(z_0, \varepsilon)} \subset B(z_0, r) \subset U$, so we have $|f(z)| = |f(z_0)|$ for all z with $|z - z_0| = \varepsilon \leq r$, and thus $|f|$ is constant on all of $\overline{B(z_0, r)} \subset U$. $\square$

Proof of Theorem 11.5.5. Assume, by way of contradiction, that $|f|$ attains its maximum at $z_0 \in U$. By Lemma 11.5.6 the modulus $|f|$ is constant in a small neighborhood around z_0 . We will show it is constant on all of U .

For any $w \in U$, if $|f(w)| < |f(z_0)|$, then choose a path γ in U from z_0 to w . Since γ is compact and U^c is closed, the distance $\varepsilon = d(\gamma, U^c) = \inf\{|c - d| \mid c \in \gamma, d \notin U\}$ is strictly positive (see Exercise 5.33). Now choose a sequence of points $z_1, \dots, z_n = w$ on γ such that $|z_i - z_{i-1}| < \varepsilon$; see Figure 11.3. Let z_k be the first such point for which $|f(z_k)| < |f(z_0)|$.

The maximum of $|f(z)|$ is $|f(z_0)| = |f(z_{k-1})|$, and it is attained at z_{k-1} , so by the lemma $|f(z)|$ is constant on $B(z_{k-1}, \varepsilon)$. Since $z_k \in B(z_{k-1}, \varepsilon)$, we have $|f(z_k)| = |f(z_{k-1})| = |f(z_0)|$, which is a contradiction.

This shows that there are no points w in U for which $|f(w)| < |f(z_0)|$; hence $|f|$ is constant on U . The theorem now follows from Proposition 11.1.7. $\square$

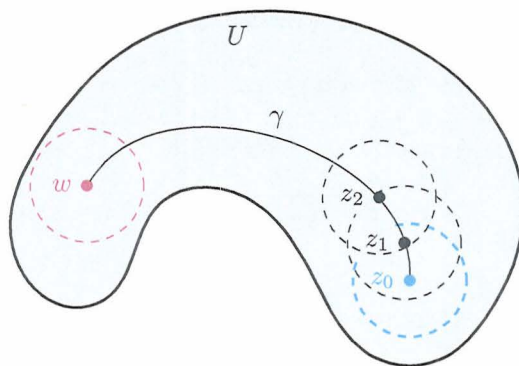


Figure 11.3. Representation of the proof of Theorem 11.5.5. If $\|f\|$ attains its maximum at z_0 (blue) but is smaller at w (red), then we can connect z_0 and w with a path γ , covered by a finite number of ε -balls inside U such that the centers z_i of the balls are less than ε apart. The previous lemma guarantees that $\|f\|$ is constant on the first ball (blue), and hence on the second, and so forth, all the way to the ball around w (red).

Corollary 11.5.7. *If f is $\mathbb{C}$ -valued, continuous on a compact set D , and holomorphic in the interior of D , then $|f|$ attains its maximum on the boundary of D .*

Proof. Since $|f|$ is continuous on a compact set, it must attain its maximum somewhere on D . If it is constant, $|f|$ attains its maximum at every point of D . If it is not constant, then the maximum modulus principle guarantees the maximum cannot be attained on the interior of D ; hence it must be on the boundary. $\square$

11.6 Power Series and Laurent Series

Power series are a very useful way to represent functions. In this section we show that *every* function that is holomorphic on an open set U can be written as a power series in a neighborhood of any point of U . Moreover, if a function has only isolated points where it fails to be holomorphic, then in a neighborhood of each of these isolated points it can be written as a power series with negative powers as well as positive powers; these series are called *Laurent series*.

11.6.1 Power Series

The power series expansion of a holomorphic function at a point z_0 is given by taking the limit of its Taylor polynomials at z_0 . A real function of one variable may be infinitely differentiable in a neighborhood of a point and yet its Taylor series at that point might not converge. But the following theorem shows that holomorphic functions are, once again, much more special than real differentiable functions because their Taylor series always converge in some open ball.

Theorem 11.6.1. *If f is holomorphic in an open neighborhood U of z_0 , then f is analytic; that is, in any open ball $B(z_0, r) \subset U$ that lies entirely in U it can be written as a power series that converges uniformly on compact subsets. Moreover, the power series expansion of f about z_0 is given by Taylor's formula:*

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z - z_0)^k. \quad (11.18)$$

The expansion (11.18) is called the Taylor expansion or the Taylor series of f at z_0 .

Proof. For any $0 < \varepsilon < r$ consider the closed ball $D = \overline{B(z_0, r - \varepsilon)} \subset B(z_0, r) \subset U$. We need only show that the series in (11.18) converges uniformly on D . Let γ be the circle $\{w \in \mathbb{C} \mid r - \varepsilon/2 = |w - z_0|\}$ with the usual, counterclockwise orientation. Note that γ lies entirely within $B(z_0, r)$ and completely encloses D .

For any $z \in D$ and $w \in \gamma$, expand $\frac{1}{w-z}$ as a power series in z_0 to get

$$\frac{1}{w-z} = \frac{1}{(w-z_0)} \frac{1}{\left(1 - \frac{z-z_0}{w-z_0}\right)} = \frac{1}{(w-z_0)} \sum_{k=0}^{\infty} \left(\frac{z-z_0}{w-z_0}\right)^k. \quad (11.19)$$

We have $|z - z_0| \leq r - \varepsilon$ and $|w - z_0| = r - \varepsilon/2$ for all $w \in \gamma$ and $z \in D$; so the equality in (11.19) holds, since $\frac{z-z_0}{w-z_0} < 1$. This implies that the series (11.19)

converges uniformly and absolutely, as a series of functions in z and w on $D \times \gamma$. Since γ is compact, $f(w)$ is bounded for $w \in \gamma$, so the series $\sum_{k=0}^{\infty} f(w) \frac{(z-z_0)^k}{(w-z_0)^{k+1}}$ is also uniformly and absolutely convergent on $D \times \gamma$.

Combining (11.19) with the Cauchy integral formula gives

$$f(z) = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(w)}{w-z} dw = \frac{1}{2\pi i} \oint_{\gamma} \sum_{k=0}^{\infty} f(w) \frac{(z-z_0)^k}{(w-z_0)^{k+1}} dw.$$

Since integration is a bounded linear operator, it commutes with uniform limits, and hence with uniformly convergent sums. Thus, we have

$$f(z) = \frac{1}{2\pi i} \sum_{k=0}^{\infty} \oint_{\gamma} f(w) \frac{(z-z_0)^k}{(w-z_0)^{k+1}} dw = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z-z_0)^k,$$

where the last line follows from Cauchy's differentiation formula (11.15).

This shows that the Taylor series for $f(z)$ converges uniformly on any closed ball $\overline{B}(z_0, r - \varepsilon)$ contained in any open ball $B(z_0, r) \subset U$. $\square$

Remark 11.6.2. We have already seen (see Theorem 11.2.8) that analytic functions are holomorphic. The previous theorem guarantees that the converse holds, so a complex function is holomorphic if and only if it is analytic. Because of this theorem, many people use the terms *holomorphic* and *analytic* interchangeably.

Proposition 11.6.3. *A convergent power series expansion*

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k$$

around z_0 is unique and is equal to the Taylor series.

Proof. By Theorem 11.2.8 we may differentiate term by term. Differentiating n times and substituting in $z = z_0$ shows that $f^{(n)}(z_0) = n!a_n$. So every expansion in $z - z_0$ that converges in a neighborhood of z_0 is equal to the Taylor series. $\square$

Corollary 11.6.4. *If f is holomorphic in a path-connected, open set $U \subset \mathbb{C}$, and if there exists some $z_0 \in U$ such that $f^{(n)}(z_0) = 0$ for all $n \in \mathbb{N}$, then $f(z) = 0$ on all of U .*

Proof. Assume, by way of contradiction, that $w \in U$ satisfies $f(w) \neq 0$. Choose a piecewise-smooth path γ from z_0 to w .

The Taylor series of f converges to f in the largest open ball $B(z_0, r)$ contained in U . Since the Taylor series is identically zero, the function f is identically zero on this ball. Just as in the proof of the maximum modulus principle, let ε be the distance from γ to U^c , choose a sequence of points $z_1, z_2, \dots, z_n = w$ along the path γ that are no more than distance $\varepsilon/2$ apart. Each ball $B(z_i, \varepsilon)$ is contained in U , and so f has a power series centered at z_i that converges in $B(z_i, \varepsilon)$.

Let z_{k+1} be the first point in the sequence for which the power series expansion of f around z_{k+1} is not identically zero. Since f is identically zero on the ball $B(z_k, \varepsilon)$, and since z_{k+1} lies in $B(z_k, \varepsilon)$, all the derivatives of f vanish at z_{k+1} , so the power series expansion of $f(z_{k+1})$ around the point z_{k+1} is also identically 0—a contradiction. $\square$

11.6.2 Zeros of Analytic Functions

Definition 11.6.5. Suppose f is analytic in an open neighborhood of z_0 with Taylor expansion

$$f(z) = a_n(z - z_0)^n + a_{n+1}(z - z_0)^{n+1} + \cdots = \sum_{k=n}^{\infty} a_k(z - z_0)^k,$$

where $a_n \neq 0$. We say that f has a zero of order n (or a zero of multiplicity n at z_0).

Proposition 11.6.6. If f is holomorphic in a neighborhood of a zero z_0 of order n , then f can be factored as $f(z) = (z - z_0)^n g(z)$, where $g(z)$ is a holomorphic function that does not vanish at z_0 . Moreover, there is an open neighborhood V around z_0 such that $f(z) \neq 0$ for every $z \in V \setminus z_0$.

Proof. It is clear from the Taylor expansion that f factors as $(z - z_0)^n g(z)$ with g analytic. Since g is continuous near z_0 , there is a neighborhood V of z_0 where $g(z) \neq 0$ if $z \in V$. Since the polynomial $(z - z_0)^n$ vanishes only at $z = z_0$, the product $(z - z_0)^n g(z) = f(z)$ cannot vanish on $V \setminus z_0$. $\square$

Corollary 11.6.7 (Local Isolation of Zeros). If f is holomorphic in a path-connected open set U , and if there is a sequence $(z_k)_{k=0}^{\infty}$ of distinct points in U converging to any $w \in U$ such that $f(z_k) = 0$ for all k , then $f(z) = 0$ for all $z \in U$.

Proof. The convergent sequence $(z_k)_{k=0}^{\infty}$ of zeros must intersect every neighborhood of w ; hence a neighborhood with no additional zeros, as described in Proposition 11.6.6, cannot exist. Thus, f must be identically zero on a neighborhood of w . By Corollary 11.6.4 f must be zero on all of U . $\square$

11.6.3 Laurent Expansions

We must also consider functions, like $1/(z - z_0)^2$ and $\cos(z)/(z - z_0)$, that are holomorphic on a punctured neighborhood of some point z_0 but are not holomorphic at z_0 .

We can no longer use Taylor series expansions to write these as a convergent power series around z_0 , since a convergent power series would be continuous at z_0 , and these functions are not continuous at z_0 . But if we allow negative powers, we can write the function as a series expansion, and much of the theory of power series still applies.

Theorem 11.6.8 (Laurent Expansion). *If f is holomorphic on the annulus $A = \{z \in \mathbb{C} \mid r < |z - z_0| < R\}$ for some $z_0 \in \mathbb{C}$ (we also allow $r = 0$ or $R = \infty$ or both), then the function f can be written in the form*

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n, \quad (11.20)$$

where we can decompose (11.20) as a sum of power series

$$f(z) = \sum_{n=0}^{\infty} c_n (z - z_0)^n + \sum_{n=1}^{\infty} c_{-n} \left(\frac{1}{z - z_0} \right)^n, \quad (11.21)$$

and both of these series converge uniformly and absolutely on every closed annulus of the form $D_{\rho, \varrho} = \{z \in \mathbb{C} \mid \rho \leq |z - z_0| \leq \varrho\}$ for $r < \rho < \varrho < R$. Furthermore, if γ is any circle around z_0 with radius $r + \varepsilon < R$, for some $\varepsilon > 0$, then for each integer n the coefficients are given by

$$c_n = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(w)}{(w - z_0)^{n+1}} dw. \quad (11.22)$$

Proof. Choose $\varepsilon_1, \varepsilon_2 > 0$ such that

$$r < r + \varepsilon_1 < \rho < \varrho < r + \varepsilon_2 < R.$$

Let γ_1 be the circle about z_0 of radius $r + \varepsilon_1$, and let γ_2 be the circle about z_0 of radius $r + \varepsilon_2$; see Figure 11.4.

By the usual change-of-path arguments, the integral (11.22) is independent of ε , provided $0 < \varepsilon < R - r$, so it is the same for ε_1 and ε_2 . For each $n \in \mathbb{Z}$, set

$$c_n = \frac{1}{2\pi i} \oint_{\gamma_1} \frac{f(w)}{(w - z_0)^{n+1}} dw = \frac{1}{2\pi i} \oint_{\gamma_2} \frac{f(w)}{(w - z_0)^{n+1}} dw.$$

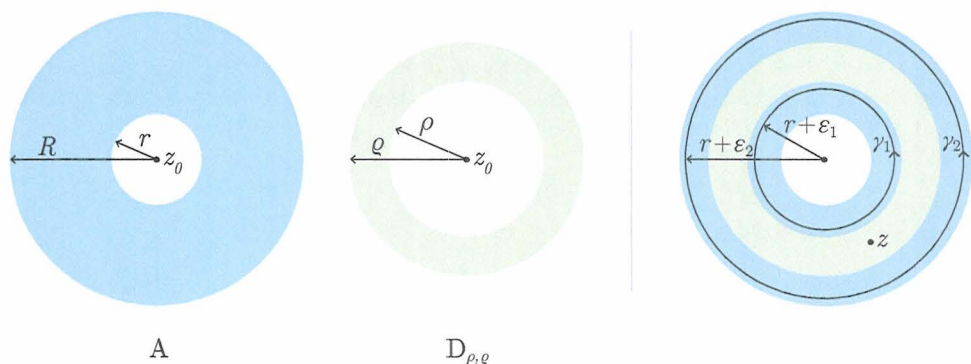


Figure 11.4. An open annulus A (blue), containing a closed annulus $D_{\rho, \varrho}$ (green), as in Theorem 11.6.8, with $r < r + \varepsilon_1 < \rho < |z| < \varrho < r + \varepsilon_2 < R$, and with circles γ_1 and γ_2 of radius $r + \varepsilon_1$ and $r + \varepsilon_2$, respectively.

By Cauchy's integral formula, for any $z \in D_{\rho, \varrho}$ we have

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_2 - \gamma_1} \frac{f(w)}{w - z} dw = \frac{1}{2\pi i} \oint_{\gamma_2} \frac{f(w)}{w - z} dw - \frac{1}{2\pi i} \oint_{\gamma_1} \frac{f(w)}{w - z} dw. \quad (11.23)$$

Expand $\frac{f(w)}{w - z}$ as in (11.19) to get

$$\frac{f(w)}{w - z} = \sum_{k=0}^{\infty} \frac{f(w)(z - z_0)^k}{(w - z_0)^{k+1}}. \quad (11.24)$$

This converges uniformly and absolutely as a function in w and z on the compact set $\gamma_2 \times D_{\rho, \varrho}$, so we may integrate the first half of (11.23) term by term to get

$$\frac{1}{2\pi i} \oint_{\gamma_2} \frac{f(w)}{w - z} = \sum_{k=0}^{\infty} \frac{1}{2\pi i} \oint_{\gamma_2} \frac{f(w)}{(w - z_0)^{k+1}} (z - z_0)^k dw = \sum_{k=0}^{\infty} c_k (z - z_0)^k,$$

and this series converges uniformly and absolutely on $D_{\rho, \varrho}$.

Alternatively, we can interchange the roles of z and w in (11.19) to get an expansion for $\frac{-1}{w - z} = \frac{1}{z - w}$ as

$$\frac{1}{z - w} = \sum_{k=0}^{\infty} \frac{(w - z_0)^k}{(z - z_0)^{k+1}},$$

which converges uniformly and absolutely as a function in w and z on $\gamma_1 \times D_{\rho, \varrho}$. Integrating the second half of (11.23) term by term gives

$$-\frac{1}{2\pi i} \oint_{\gamma_1} \frac{f(w)}{w - z} dw = \sum_{k=0}^{\infty} \frac{c_{-k}}{(z - z_0)^k}.$$

To see that this second series converges uniformly and absolutely on $D_{\rho, \varrho}$ substitute $t = 1/(z - z_0)$ and use the previous results for power series in t . $\square$

Proposition 11.6.9. *For a given choice of $A = \{z \in \mathbb{C} \mid r < |z - z_0| < R\}$, the Laurent expansion (11.20) of f is unique.*

Proof. The proof is similar to that for Taylor series, but instead of differentiating term by term, write the expansion of $\frac{f(z)}{(z - z_0)^{k+1}}$ and integrate term by term around a contour γ centered at z_0 with radius $r + \varepsilon$ in A , using Lemma 11.3.5. The details are Exercise 11.28. $\square$

Remark 11.6.10. Computing the Laurent expansion of a function is often much more difficult than computing Taylor expansions. But in many cases finding just a few terms of the Laurent expansion is sufficient to solve the problem at hand, and often this can be done without much difficulty.

Example 11.6.11. To find the Laurent expansion of $\cos(z)/z^3$ around $z_0 = 0$ we can expand $\cos(z)$ as $\cos(z) = \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k}}{(2k)!}$ and divide by z^3 to get

$$\cos(z)/z^3 = \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k-3}}{(2k)!} = \frac{1}{z^3} - \frac{1}{2!z} + \frac{z}{4!} + \cdots.$$

Example 11.6.12. To compute the Laurent expansion of $\frac{z}{z^2+1}$ around $z_0 = i$, solve for the partial fraction decomposition $\frac{z}{z^2+1} = \frac{a}{z+i} + \frac{b}{z-i}$ to get $a = 1/2 = b$, so $\frac{z}{z^2+1} = \frac{1/2}{z+i} + \frac{1/2}{z-i}$. Around $z_0 = i$ the function $\frac{1/2}{z-i}$ is holomorphic, and expanding it as a power series in $z-i$ gives

$$\frac{1/2}{z+i} = \frac{1/2}{2i + (z-i)} = \frac{1/4i}{1 - i(z-i)/2} = \frac{1}{4i} \sum_{k=0}^{\infty} (i/2)^k (z-i)^k.$$

This converges when $|(i/2)(z-i)| < 1$ and diverges at $(i/2)(z-i) = 1$, so the radius of convergence is 1. Thus we have

$$\frac{z}{z^2+1} = \frac{1/2}{z-i} + \frac{1}{4i} \sum_{k=0}^{\infty} (i/2)^k (z-i)^k.$$

Since $\frac{1/2}{z-i}$ is holomorphic away from $z = i$, the Laurent expansion is valid in the punctured ball $B(i, 1) \setminus \{i\}$.

11.7 The Residue Theorem

When a function is holomorphic on a simply connected set U , the Cauchy–Goursat theorem guarantees that the value of its contour integral on any simple closed contour Γ is zero. The residue theorem gives a way to calculate the contour integral when the function fails to be holomorphic at a finite number of points enclosed by the contour but is holomorphic on the contour. Before we can understand this theorem, we must first understand the sorts of points that occur where the function is not holomorphic.

11.7.1 Isolated Singularities

Definition 11.7.1. If f is holomorphic on the set $\{z \in \mathbb{C} \mid 0 < |z - z_0| < R\}$ and not holomorphic at z_0 , we say that z_0 is an isolated singularity of f . Let z_0 be an isolated singularity of f , and let $\sum_{k=-\infty}^{\infty} c_k(z - z_0)^k$ be the Laurent expansion on a small punctured neighborhood $B(z_0, r) \setminus \{z_0\}$ of z_0 .

- (i) The series $\sum_{k=-\infty}^{-1} c_k(z - z_0)^k$ is the principal part of the Laurent series for f at z_0 .

- (ii) If the coefficients c_k all vanish for $k < 0$ (if the principal part is zero), then the singularity is called removable. In this case f can be extended to a holomorphic function on $B(z_0, r)$ using the power series $f(z) = \sum_{k=0}^{\infty} c_k(z - z_0)^k$.
- (iii) If the principal part has only a finite number of nonzero terms, so that $f = \sum_{k=-N}^{\infty} c_k(z - z_0)^k$, with $c_{-N} \neq 0$, then z_0 is called a pole of f of order N .
- (iv) A pole of first order, meaning that $c_k = 0$ for all $k < -1$, and $c_{-1} \neq 0$, is called a simple pole.
- (v) If the principal part has an infinite number of nonzero terms, then z_0 is an essential singularity of f .

Example 11.7.2.

- (i) $\frac{\sin z}{z}$ has a removable singularity at $z_0 = 0$, since

$$\frac{\sin z}{z} = \frac{1}{z} \left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots \right) = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \cdots.$$

- (ii) $\frac{e^z}{z^2}$ has a pole of order 2 at $z_0 = 0$ since $\frac{e^z}{z^2} = \frac{1}{z^2} \left(1 + z + \frac{z^2}{2!} + \cdots \right)$.

- (iii) $e^{1/z} = \sum_{k=-\infty}^0 z^k / (-k)!$ has an essential singularity at $z = 0$.

Example 11.7.3. Assume that f and g are both holomorphic and $\mathbb{C}$ -valued in a neighborhood of z_0 , that f has a zero at z_0 of order k , and that g has a zero of order ℓ at z_0 . We can write $f(z) = (z - z_0)^k F(z)$ and $g(z) = (z - z_0)^\ell G(z)$, where F and G are both holomorphic and do not vanish at z_0 .

- (i) If $k \geq \ell$, then f/g is undefined at z_0 , but $f/g = (z - z_0)^{k-\ell} F/G$ away from z_0 , and F/G and $(z - z_0)^{k-\ell}$ are both holomorphic at z_0 , so the singularity at z_0 is removable.
- (ii) If $k < \ell$, then f/g has a pole of order $\ell - k$ at z_0 .

Definition 11.7.4. A function that is holomorphic in an open set U , except for poles in U , is meromorphic in U . If U is not specified, it is assumed to be all of $\mathbb{C}$.

Example 11.7.5. Rational functions are functions of the form $p(z)/q(z)$, where $p, q \in \mathbb{C}[z]$. If p and q have no common zeros and q is not identically zero, then these are meromorphic on $\mathbb{C}$, since their only singularities are isolated and of finite order. If p and q do have common zeros, then dividing out all common factors gives a new rational function a/b that is meromorphic

and that agrees with p/q at all points of $\mathbb{C}$ except at the zeros of q . It is standard practice to replace p/q by a/b everywhere but still write p/q to denote the function a/b . We will also do this everywhere without further comment.

11.7.2 Residues and Winding Numbers

Although the computation of the full Laurent series for a function may be difficult, in many cases it turns out that the only thing we really need from the Laurent series is the coefficient of $(z - z_0)^{-1}$. The main reason for this is the computation we did in Lemma 11.3.5, which says that the integral of $(z - z_0)^k$ around a simple closed curve γ containing z_0 vanishes unless $k = -1$.

Definition 11.7.6. Let f be holomorphic on a punctured ball $B(z_0, r) \setminus \{z_0\} = \{z \in \mathbb{C} \mid 0 < |z - z_0| < R\}$, and let γ be a simple closed contour in $B(z_0, r) \setminus \{z_0\}$ with z_0 in the interior of γ . The number

$$\frac{1}{2\pi i} \oint_{\gamma} f(z) dz$$

is called the residue of f at z_0 and is denoted $\text{Res}(f, z_0)$.

Proposition 11.7.7. Let f be holomorphic on a punctured ball $B(z_0, r) \setminus \{z_0\} = \{z \in \mathbb{C} \mid 0 < |z - z_0| < r\}$, with the expansion (11.20). The residue is given by

$$\text{Res}(f, z_0) = c_{-1}.$$

Proof. This follows immediately from the fact that the Laurent expansion converges uniformly on compact subsets, so integration commutes with the sums and we get

$$\begin{aligned} \text{Res}(f, z_0) &= \frac{1}{2\pi i} \oint_{\gamma} f(z) dz = \frac{1}{2\pi i} \oint_{\gamma} \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n dz \\ &= \frac{1}{2\pi i} \sum_{n=-\infty}^{\infty} \oint_{\gamma} c_n (z - z_0)^n dz = \frac{1}{2\pi i} \oint_{\gamma} \frac{c_{-1}}{z - z_0} dz = c_{-1}. \quad \square \end{aligned}$$

Proposition 11.7.8. If f has an isolated singularity at z_0 , then the following hold:

- (i) The singularity at z_0 is removable if and only if $\lim_{z \rightarrow z_0} f(z)$ is finite.
- (ii) If $\lim_{z \rightarrow z_0} (z - z_0)^k f(z)$ exists (is finite) for some $k \geq 0$, then the singularity is removable or is a pole of order less than or equal to k .
- (iii) If $\lim_{z \rightarrow z_0} (z - z_0)f(z)$ exists, then it is equal to the residue:

$$\text{Res}(f, z_0) = \lim_{z \rightarrow z_0} (z - z_0)f(z). \quad (11.25)$$

Proof. This follows from the Laurent expansion; see Exercise 11.29. $\square$

If the contour γ is closed but not simple, or if it does not contain z_0 , then the integral of f around γ depends not only on the coefficient c_{-1} (the residue), but also on the contour itself.

If γ is a simple closed curve traversed once in a counterclockwise direction with z_0 in the interior of γ , then the integral of $1/(z - z_0)$ around γ is $2\pi i$. A little thought shows that if γ circles around z_0 a total of k times in a counterclockwise direction, then the integral is $2\pi ki$. If γ does not enclose the point z_0 at all, there exists a simply connected region containing γ but not z_0 where $\frac{1}{z - z_0}$ is holomorphic. When the function $\frac{1}{z - z_0}$ is holomorphic on a simply connected region containing γ , the Cauchy–Goursat theorem guarantees that the integral over this contour is 0. These observations motivate the following definition.

Definition 11.7.9. Let γ be a closed contour on $\mathbb{C}$ and $z_0 \in \mathbb{C}$ a point not on the contour γ . The winding number of γ with respect to z_0 is

$$I(\gamma, z_0) = \frac{1}{2\pi i} \oint_{\gamma} \frac{dz}{z - z_0}. \quad (11.26)$$

Nota Bene 11.7.10. The winding number essentially counts the total number of times a closed curve travels counterclockwise around a given point.

Example 11.7.11.

- (i) Lemma 11.3.5 and Example 11.3.16 show that for any simple closed contour σ the winding number $I(\sigma, 0)$ is 1 if z_0 is contained in σ and is zero otherwise.
- (ii) For the curve γ in Figure 11.5 we have

$$I(\gamma, z_1) = 1, \quad I(\gamma, z_2) = 2, \quad \text{and} \quad I(\gamma, z_3) = 0.$$

The next lemma is a straightforward consequence of the previous definitions.

Lemma 11.7.12. Let U be a simply connected open set, and let γ be a closed contour in U . If $N(z) = \sum_{k=0}^{\infty} \frac{b_k}{(z - z_0)^k}$ is uniformly convergent on compact subsets in the punctured set $U \setminus \{z_0\}$, then we have

$$\frac{1}{2\pi i} \oint_{\gamma} N(z) dz = \text{Res}(N, z_0) I(\gamma, z_0).$$

Proof. The proof is Exercise 11.30. $\square$

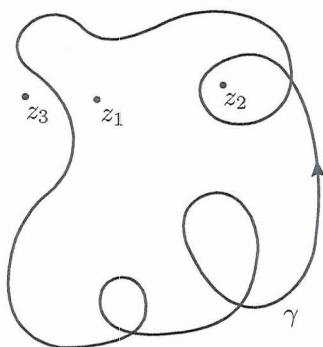


Figure 11.5. Examples of winding numbers as described in Definition 11.7.9. In this example $I(\gamma, z_1) = 1$, whereas $I(\gamma, z_2) = 2$, and $I(\gamma, z_3) = 0$.

11.7.3 The Residue Theorem

We are now ready for the main theorem of this section.

Theorem 11.7.13 (Residue Theorem). Let U be a simply connected open set, and let f be holomorphic on all of U , except for a finite number of isolated singularities $\{z_1, \dots, z_n\}$; that is, f is holomorphic on $U \setminus \{z_1, \dots, z_n\}$. Assume γ is a closed curve in U and that no z_i lies on γ . We have

$$\oint_{\gamma} f(z) dz = 2\pi i \sum_{i=1}^n \text{Res}(f, z_i) I(\gamma, z_i). \quad (11.27)$$

Remark 11.7.14. The idea behind this theorem should be fairly clear: use the usual change-of-path method to replace the contour γ with a sum of little circles $\sum_{j=1}^n \gamma_j$, one around each isolated singularity z_j , traversed $I(\gamma, z_j)$ times. To compute the integral $\oint_{\gamma_j} f(z) dz$, just use the Laurent expansion and integrate term by term to get the desired result.

We need to be more careful, however, since there is some sloppiness in the claim about winding numbers matching the number of times the path happens to encircle a given singularity. For the careful proof, it turns out to be easier to use a slightly different approach.

Proof. For each singularity z_j expand f as a convergent Laurent series on a punctured disk around z_j ,

$$f(z) = \sum_{m=-\infty}^{\infty} c_m^{(j)} (z - z_j)^m,$$

and consider the principal parts

$$N_j(z) = \sum_{m=-\infty}^{-1} c_m^{(j)} (z - z_j)^m = \sum_{k=1}^{\infty} \frac{c_{-k}^{(j)}}{(z - z_j)^k}.$$

Each N_j converges uniformly on compact subsets of $\mathbb{C} \setminus \{z_j\}$ and hence is holomorphic on $U \setminus \{z_j\}$.

Let $g(z)$ be the function obtained by subtracting from f the sum of all the principal parts:

$$g(z) = f(z) - \sum_{j=1}^n N_j(z).$$

Near z_j the positive part of the Laurent expansion of f is of the form $\sum_{m=0}^{\infty} c_m^{(j)}(z - z_j)^m$, which converges uniformly on compact subsets and defines a holomorphic function on a neighborhood B_j of z_j . In B_j , the principal parts $N_\ell(z)$ are all holomorphic if $\ell \neq j$, so the function $G(z) = \sum_{m=0}^{\infty} c_m(z - z_j)^m - \sum_{\ell \neq j} N_\ell$ is holomorphic on B_j . Since $g(z) = G(z)$ at every point of $B_j \setminus \{z_j\}$, the function $g(z)$ has a removable singularity at $z = z_j$.

Since this holds for every $j \in \{1, \dots, n\}$, the Cauchy–Goursat theorem gives

$$0 = \oint_{\gamma} g(z) dz = \oint_{\gamma} f(z) dz - \sum_{j=1}^n \oint_{\gamma} N_j(z) dz.$$

Thus, we see

$$\begin{aligned} \oint_{\gamma} f(z) dz &= \sum_{j=1}^n \oint_{\gamma} N_j(z) dz \\ &= 2\pi i \sum_{j=1}^n \text{Res}(f, z_j) I(\gamma, z_j), \end{aligned}$$

where the last equality follows from Lemma 11.7.12. $\square$

The residue theorem is very powerful, especially when combined with another method for calculating the residues easily. The next proposition gives one method for doing this.

Proposition 11.7.15. *Assume h is a $\mathbb{C}$ -valued function and g and h are both holomorphic in a neighborhood of z_0 . If $g(z_0) \neq 0$, $h(z_0) = 0$, and $h'(z_0) \neq 0$, then the function $g(z)/h(z)$ has a simple pole at z_0 and*

$$\text{Res}\left(\frac{g(z)}{h(z)}, z_0\right) = \frac{g(z_0)}{h'(z_0)}. \quad (11.28)$$

Proof. Note that

$$\lim_{z \rightarrow z_0} \frac{h(z) - h(z_0)}{z - z_0} = \lim_{z \rightarrow z_0} \frac{h(z)}{z - z_0} = h'(z_0) \neq 0.$$

Hence

$$\lim_{z \rightarrow z_0} \frac{z - z_0}{h(z)} = \frac{1}{h'(z_0)}$$

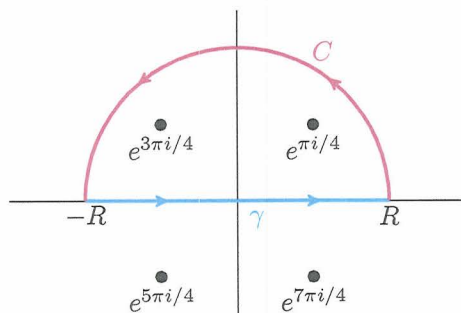


Figure 11.6. The D-shaped contour of radius R , as described in Example 11.7.17. The contour D consists of a half circle C (red) and a line segment γ (blue) and contains two of the four poles of $1/x^4 + 1$. It is not hard to show that as $R \rightarrow \infty$ the integral along C goes to 0. Thus, in the limit, the integral along γ is the same as the integral over D , which can be computed using the residue theorem.

exists, and we get

$$\frac{g(z_0)}{h'(z_0)} = \lim_{z \rightarrow z_0} (z - z_0) \frac{g(z)}{h(z)} = \text{Res} \left(\frac{g}{h}, z_0 \right),$$

where the last equality follows from Proposition 11.7.8. $\square$

Example 11.7.16. Let γ be a circle centered at $z_0 = 0$ with radius 2. From Proposition 11.7.15 we have

$$\begin{aligned} \int_{\gamma} \frac{dz}{z^2 - 1} &= 2\pi i \left[\text{Res} \left(\frac{1}{z^2 - 1}, 1 \right) + \text{Res} \left(\frac{1}{z^2 - 1}, -1 \right) \right] \\ &= 2\pi i \left[\frac{1}{2(1)} + \frac{1}{2(-1)} \right] = 0. \end{aligned}$$

Example 11.7.17. Contour-integral techniques are useful for computing real integrals that would otherwise be extremely difficult to evaluate. For example, consider the integral $\int_{-\infty}^{\infty} \frac{dx}{x^4 + 1}$. It is not difficult to verify that this improper integral converges and that it can be computed using the symmetric limit

$$\int_{-\infty}^{\infty} \frac{dx}{x^4 + 1} = \lim_{R \rightarrow \infty} \int_{-R}^R \frac{dx}{x^4 + 1}.$$

To compute this integral, consider the contour integral over the following D-shaped contour. Let C be the upper half circle of radius $R > 2$ centered at the origin, as shown in Figure 11.6, let γ be the line segment from $(-R, 0)$ to $(R, 0)$, and let $D = C + \gamma$ be their sum. We have

$$\oint_D \frac{dz}{z^4 + 1} = \int_C \frac{dz}{z^4 + 1} + \int_\gamma \frac{dx}{x^4 + 1} = \int_C \frac{dz}{z^4 + 1} + \int_{-R}^R \frac{dx}{x^4 + 1}.$$

From Proposition 11.7.15 we have

$$\begin{aligned} \oint_D \frac{dz}{z^4 + 1} &= 2\pi i \left[\operatorname{Res} \left(\frac{1}{z^4 + 1}, e^{\pi i/4} \right) + \operatorname{Res} \left(\frac{1}{z^4 + 1}, e^{3\pi i/4} \right) \right] \\ &= 2\pi i \left[\frac{1}{4(e^{\pi i/4})^3} + \frac{1}{4(e^{3\pi i/4})^3} \right] = -\frac{\pi}{2i} \left[e^{-3i\pi/4} + e^{-9\pi i/4} \right] \\ &= \frac{\pi}{\sqrt{2}}. \end{aligned}$$

Now consider the integral just over C . Parametrizing C by Re^{it} with $t \in [0, \pi]$ gives

$$\begin{aligned} \left| \int_C \frac{dz}{z^4 + 1} \right| &= \left| \int_0^\pi \frac{Rie^{it} dt}{R^4 e^{4it} + 1} \right| \\ &\leq \int_0^\pi \left| \frac{Rie^{it}}{R^4 e^{4it} + 1} \right| dt \leq \int_0^\pi \frac{R}{R^4 - 1} dt \rightarrow 0 \quad \text{as } R \rightarrow \infty. \end{aligned}$$

Therefore we have

$$\int_{-\infty}^{\infty} \frac{dx}{x^4 + 1} = \lim_{R \rightarrow \infty} \left(\int_{-R}^R \frac{dx}{x^4 + 1} + \int_C \frac{dz}{z^4 + 1} \right) = \lim_{R \rightarrow \infty} \oint_D \frac{dz}{z^4 + 1} = \frac{\pi}{\sqrt{2}}.$$

11.8 * The Argument Principle and Its Consequences

In this section we use the residue theorem to derive three important results that have many practical applications. The first is the *argument principle*, which is useful to determine the number of zeros and poles a given meromorphic function has. The second result, *Rouché's theorem*, is useful for determining the locations of zeros of a complicated analytic function when the location of zeros of a simpler analytic function is known. Finally, the third is the *holomorphic open mapping theorem*, which states that a nonconstant holomorphic function must map an open set to an open set.

11.8.1 The Argument Principle

The name “argument principle” requires some explanation. The *argument* $\arg(z)$ of a point $z = re^{i\theta}$ is the angle θ between the positive real axis to the vector representing z . This is not a well-defined function on $\mathbb{C}$ because $e^{i\theta} = e^{i(\theta+2\pi k)}$ for any integer k .

Although the *value* of the argument at a point may not be well defined, it does make sense to consider the *change in argument* as z traces a closed contour σ . In fact this is exactly what the winding number $I(\sigma, 0)$ is tracking. That is to say,

$2\pi I(\sigma, 0)$ is the total change in angle (argument) for the contour σ . If $\sigma = f \circ \gamma$, then $2\pi I(f \circ \gamma, 0)$ is the total change in argument for $f \circ \gamma$, at least if $f(w) \neq 0$ for every $w \in \gamma$.

Theorem 11.8.1 (Zero and Pole Counting Formula). *Let U be a simply connected open subset of $\mathbb{C}$, and let γ be a closed contour in U . Let f be a $\mathbb{C}$ -valued meromorphic function on U , with poles $\{w_1, \dots, w_n\}$, and zeros $\{z_1, \dots, z_m\}$, none of which lies on the contour γ . If $b_1, \dots, b_n$ are the respective orders of the poles and $a_1, \dots, a_m$ are the respective multiplicities of the zeros, then*

$$\oint_{\gamma} \frac{f'(z)}{f(z)} dz = 2\pi i \left(\sum_{k=1}^m a_k I(\gamma, z_k) - \sum_{k=1}^n b_k I(\gamma, w_k) \right). \quad (11.29)$$

Proof. Because f is meromorphic in U , the Laurent series for f at any point z_0 in U has the form

$$f(z) = \sum_{j=k}^{\infty} c_j (z - z_0)^j$$

for some (finite) integer k . Moreover, we may assume that $c_k \neq 0$. This implies that we can factor f as

$$f(z) = (z - z_0)^k g(z),$$

where g is holomorphic and nonvanishing at z_0 . If $k > 0$, then f has a zero of order k at z_0 , whereas if $k < 0$, then f has a pole of order $-k$.

Setting $F(z) = f'(z)/f(z)$ we get

$$\begin{aligned} F(z) &= \frac{f'(z)}{f(z)} = \frac{k(z - z_0)^{k-1}g(z) + (z - z_0)^k g'(z)}{(z - z_0)^k g(z)} \\ &= \frac{k g(z) + (z - z_0) g'(z)}{(z - z_0) g(z)} \\ &= \frac{k}{(z - z_0)} + \frac{g'(z)}{g(z)}. \end{aligned}$$

Since g is holomorphic and nonvanishing at z_0 , the term $\frac{g'(z)}{g(z)}$ is also holomorphic near z_0 . This shows that the function $F(z)$ has residue k at z_0 . The residue theorem now gives (11.29), as desired. $\square$

Remark 11.8.2. This theorem says the integral (11.29) is always an *integer* multiple of $2\pi i$. That means we can compute it exactly by just using an approximation that is good enough to identify which multiple of $2\pi i$ it must be. If f is holomorphic, a rough numerical approximation that is just good enough to identify the nearest integer multiple of $2\pi i$ tells the exact number of zeros of f (with multiplicity) that lie inside of γ . If f is meromorphic, then such a numerical approximation tells the exact difference between the number of zeros and poles inside of γ .

Corollary 11.8.3. *Let U be a simply connected, open subset of $\mathbb{C}$, and let γ be a simple closed contour in U . Fix $w \in \mathbb{C}$, and let f be a $\mathbb{C}$ -valued holomorphic on U ,*

with $f(z) \neq w$ for every $z \in \gamma$. If N is the number of solutions of $f(z) = w$ (with multiplicities) that lie within γ , then

$$\frac{1}{2\pi i} \oint_{\gamma} \frac{f'(z)}{f(z) - w} dz = N.$$

Proof. Let $g = f - w$ and apply the zero and pole counting formula (11.29) to g on γ . $\square$

Corollary 11.8.4 (Argument Principle). *Let U be a simply connected, open subset of $\mathbb{C}$, and let γ be a simple closed contour in U . Let f be a $\mathbb{C}$ -valued function, meromorphic on U , with no zeros or poles on the contour γ , with Z zeros (counted with multiplicity) inside of γ , and with P poles (also counted with multiplicity) lying inside of γ . Define a new closed contour $f \circ \gamma$ as follows. If γ is parametrized by $z(t)$ with $t \in [a, b]$, parametrize a new contour by $w(t) = f(z(t))$, also with $t \in [a, b]$ (or if γ is piecewise smooth, construct $f \circ \gamma$ from each of the smooth pieces $w(t) = f(z(t))$ in the obvious way). In this case, we have*

$$I(f \circ \gamma, 0) = Z - P.$$

Proof. Since γ is a simple closed contour, the winding number $I(\gamma, z_0)$ is 1 for every z_0 inside of γ and 0 otherwise. We have

$$I(f \circ \gamma, 0) = \frac{1}{2\pi i} \oint_{f \circ \gamma} \frac{dw}{w} = \frac{1}{2\pi i} \oint_{\gamma} \frac{f'(z) dz}{f(z)} = Z - P,$$

where the last line follows from the zero and pole counting formula (11.29). $\square$

Example 11.8.5. Suppose

$$f(z) = \frac{(z-7)^3 z^2}{(z-6)^3 (z+2)^5 (z-1)^2}.$$

Let's evaluate the integral of f'/f around the circle of radius 4, centered at zero, traversed once, oriented counterclockwise. By Theorem 11.8.1 we have

$$\begin{aligned} \oint_{\gamma} \frac{f'(z)}{f(z)} dz &= 2\pi i \left(\sum_{k=1}^m a_k I(\gamma, z_k) - \sum_{k=1}^n b_k I(\gamma, w_k) \right) \\ &= 2\pi i (2 - (2 + 5)) = -10\pi i. \end{aligned}$$

The number of zeros inside γ is 2, and the number of poles inside γ is $2 + 5$.

11.8.2 Rouché's Theorem

Like the argument principle, Rouché's theorem is also concerned with the number of zeros of a function, but it approaches this problem from a different angle. Suppose we know that a function $f(z)$ is holomorphic on a contour γ and inside γ . If we perturb $f(z)$ by some other holomorphic function, Rouché's theorem tells us about the number of zeros of the perturbed function inside of γ .

Theorem 11.8.6 (Rouché's Theorem). *Let U be a simply connected open subset of $\mathbb{C}$, and let f and g be $\mathbb{C}$ -valued holomorphic functions on U . If γ is a simple closed contour in U , and if $|f(z)| > |f(z) - g(z)|$ for every $z \in \gamma$, then f and g have the same number of zeros, counted with multiplicity, inside of γ .*

Proof. Consider the function $F(z) = g(z)/f(z)$. The difference between the number Z of zeros and the number P of poles of F is precisely the difference between the number of zeros of g and the number of zeros of f . We show that this difference is zero.

If f or g has a zero at some $z \in \gamma$, then the hypothesis $|f(z)| > |f(z) - g(z)|$ could not hold. Therefore, the function F has no poles or zeros on γ .

For all $z \in \gamma$ we have

$$|1 - F(z)| = \left| \frac{f(z)}{f(z)} - \frac{g(z)}{f(z)} \right| = \frac{|f(z) - g(z)|}{|f(z)|} < 1.$$

Therefore, the distance from the contour $F \circ \gamma$ to 1 is always less than 1, and hence 0 does not lie within the contour $F \circ \gamma$ and $I(F \circ \gamma, 0) = 0$. This gives

$$0 = I(F \circ \gamma, 0) = Z - P. \quad \square$$

Remark 11.8.7. Sometimes Rouché's theorem is referred to as the “dog-walking” theorem. If you have ever tried to walk a dog on a leash near a lamppost or a tree you have seen this theorem in action. If the leash is short enough, you and the dog walk around the lamppost an equal number of times. But if the leash is too long, the dog may circle the lamppost more or fewer times than you do (and the leash becomes tangled).

In relation to Rouché's theorem, $f(z)$ is your path, $g(z)$ is the dog's path, and the origin is the lamppost. The maximum difference $|f(z) - g(z)|$ is the length of the leash. If the leash never extends to the lamppost, then $|f(z)| > |f(z) - g(z)|$, and the dog must circle the lamppost the same number of times as you do; see Figure 11.7.

Example 11.8.8.

- (i) To find the number of zeros of $z^5 + 8z + 10$ inside the unit circle $|z| = 1$, choose $g(z) = z^5 + 8z + 10$ and $f(z) = 10$. On the unit circle we have $|g(z) - f(z)| = |z^5 + 8z| \leq 9 < |f(z)| = 10$, so Rouché's theorem says the number of zeros of g inside the unit circle is the same as f , that is, none.

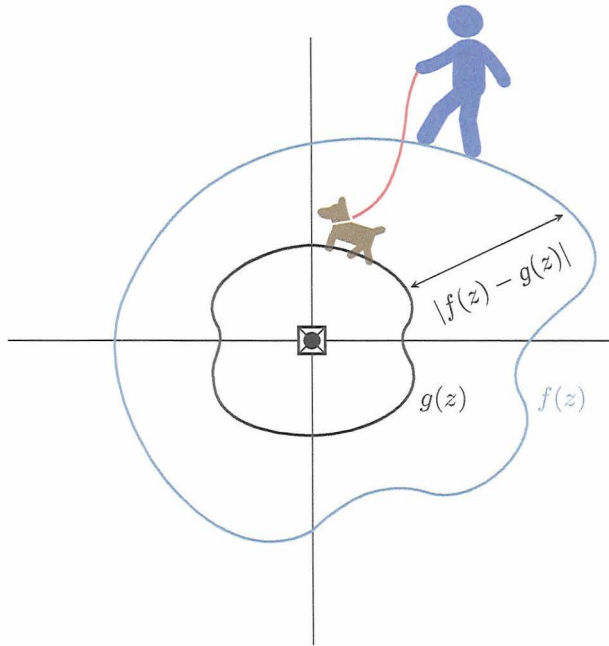


Figure 11.7. *Rouché's dog walking.* Your path (blue) corresponds to the contour $f(z)$ and the dog's path (brown) corresponds to the contour $g(z)$. At the origin is a lamppost. If the leash is not long enough to extend to the lamppost from any point of your path, then you and the dog must circle the lamppost the same number of times; see Remark 11.8.7.

- (ii) To find the number of zeros of $z^5 + 8z + 10$ inside the circle $|z| = 2$, choose $g(z) = z^5 + 8z + 10$ and $f(z) = z^5$. On the circle we have $|g(z) - f(z)| = |8z + 10| \leq 26 < |f(z)| = 32$, so Rouché's theorem says the number of zeros of g inside this circle is the same as f , that is, 5. Combining this with the previous result shows that all the zeros of g lie in the annulus $\{z \in \mathbb{C} \mid 1 < |z| < 2\}$.
- (iii) We show that for every $R > \sqrt{5}$ the function $g(z) = e^z - z^2 + 4$ has exactly 1 zero in the left half semicircle^a $S = \{z \in \mathbb{C} \mid \Re(z) \leq 0, |z| = R\} \cup \{z = iy \in \mathbb{C} \mid y \in [-R, R]\}$. Let $f(z) = (z^2 - 4)$. Since $R^2 > 5$, on the circular part $\{z \in \mathbb{C} \mid \Re(z) \leq 0, |z| = R\}$, we have

$$|f(z)| \geq R^2 - 4 > 1 \geq |e^{\Re(z)} e^{i\Im(z)}| = |e^z| = |g(z) - f(z)|.$$

On the vertical line $\{z = iy \in \mathbb{C} \mid y \in [-R, R]\}$ we have

$$|f(z)| = |-y^2 - 4| \geq 4 > 1 = |e^{iy}| = |g(z) - f(z)|.$$

So by Rouché's theorem, g has the same number of zeros as $f = (z - 2)(z + 2)$ in the semicircle S , namely, one zero.

^aRecall that $\Re(z)$ refers to the real part of z and $\Im(z)$ refers to the imaginary part of z .

Nota Bene 11.8.9. To use Rouché's theorem, you need to find a suitable function $f(z)$. A good rule of thumb is to consider a function that shares part of the function $g(z)$ but whose zeros are easier to find. For example, in part (i) of the previous example, both $f(z) = z^5$ and $f(z) = 10$ are functions for which it is easy to find the zeros. But you can check that $f(z) = z^5$ does not satisfy the hypothesis of the theorem, whereas $f(z) = 10$ does.

If you happen to choose the wrong function $f(z)$ the first time, just try another.

11.8.3 The Holomorphic Open Mapping Theorem

Recall from Theorem 5.2.3 that continuous functions pull open sets back to open sets, but for continuous functions the image of an open set need not be open. The holomorphic open mapping theorem says that for a nonconstant holomorphic function the image of an open set is always open.

Theorem 11.8.10 (Holomorphic Open Mapping Theorem). *If $U \subset \mathbb{C}$ is open, and if $f: U \rightarrow \mathbb{C}$ is a nonconstant holomorphic function, then the image $f(U)$ is open in $\mathbb{C}$.*

Proof. Given any $w_0 \in f(U)$, there is, by definition, a $z_0 \in U$ such that $w_0 = f(z_0)$. We must show that there is an $\varepsilon > 0$ such that $B(w_0, \varepsilon) \subset f(U)$.

Consider the function $g(z) = f(z) - w_0$, which is also holomorphic on U and not constant. It has a zero at z_0 . Since zeros of nonconstant holomorphic functions are isolated (see Proposition 11.6.6), there is an open neighborhood $V \subset U$ such that the only zero of $g(z)$ in V is z_0 . Choose $\delta > 0$ such that the closed ball $\overline{B}(z_0, \delta)$ lies entirely inside V .

The circle $C_\delta = \{z \in \mathbb{C} : |z - z_0| = \delta\}$ is compact, and $g(z) \neq 0$ on the circle. Let $\varepsilon > 0$ be the minimum value of $\|g(z)\|$ on C_δ . For any $w \in B(w_0, \varepsilon)$, let $h_w(z) = f(z) - w$. For any $z \in C_\delta$, we have

$$\|g(z)\| \geq \varepsilon > \|w_0 - w\| = \|g(z) - h_w(z)\|.$$

Therefore, Rouché's theorem guarantees $g(z)$ and $h_w(z)$ have the same number of zeros inside $B(z_0, \delta)$. This implies that for every $w \in B(w_0, \varepsilon)$ there is a point $z_w \in B(z_0, \delta)$ such that $0 = h_w(z_w) = f(z_w) - w$, and thus $f(z_w) = w$. That is to say, $B(w_0, \varepsilon) \subset f(B(z_0, \delta)) \subset f(U)$. $\square$

Remark 11.8.11. The holomorphic open mapping theorem shows, among other things, that no holomorphic function can map an open set to the real line, since the real line is not open in $\mathbb{C}$.

In particular, the maps $\Re(z)$, $\Im(z)$, and $|\cdot|$ cannot be holomorphic, since they all map $\mathbb{C}$ to the real line. This also shows that $\bar{z}$ cannot be holomorphic, since $2\Re(z) = z + \bar{z}$.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

11.1. Identify (and prove) the largest open subset of $\mathbb{C}$ where the following functions are holomorphic:

- (i) $g(z) = |z|$.
- (ii) $f(z) = z \sin(z)$.
- (iii) $h(z) = e^{|z|^2}$.
- (iv) $k(z) = \frac{1+z^2}{1-z^2}$.

11.2. Let

$$f(x, y) = \begin{cases} \frac{x^{4/3}y^{5/3} + ix^{5/3}y^{4/3}}{x^2 + y^2} & \text{if } z \neq 0, \\ 0 & \text{if } z = 0. \end{cases}$$

Show that the Cauchy–Riemann equations hold at $z = 0$ but that f is not holomorphic at this point. Why doesn't this contradict Theorem 11.1.4?

11.3. Let $u = e^{-x}(x \sin y - y \cos y)$. Find a v such that $f(z) = u + iv$ is holomorphic.

11.4. Prove that in polar form the Cauchy–Riemann equations can be written

$$\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r} \frac{\partial u}{\partial \theta}.$$

11.5. Prove Proposition 11.1.7. Hint: Write $\bar{f} = |f|^2/f$.

-
- 11.6. Prove Theorem 11.2.2. (Hint: For (iv) first show that any monomial az^n is holomorphic by induction on n . Then show $\sum_{j=0}^k a_j z^j$ is holomorphic by induction on k .)
- 11.7. Can a power series $\sum_{k=0}^{\infty} a_k (z-2)^k$ converge at $z=0$ but diverge at $z=3$? Prove or disprove.
- 11.8. Consider the series $(1+z)^{-1} = \sum_{k=0}^{\infty} (-z)^k$.
- (i) What is the radius of convergence R ?
 - (ii) Find a point on the boundary $|z|=R$ where the series diverges.
 - (iii) Find a convergent power series expansion for $f(z) = \log(1+z)$ around $z=0$ by integrating the series for $(1+z)^{-1}$ term by term.
 - (iv) What is the radius of convergence of this new series?
- 11.9. Let f be holomorphic on an open, connected subset A of $\mathbb{C}$. Define $g(z) = \overline{f(\bar{z})}$ on the set

$$\bar{A} = \{z \mid \bar{z} \in A\}.$$

Prove that g is holomorphic on $\bar{A}$, and that $g'(z) = \overline{f'(\bar{z})}$.

- 11.10. Show that the length of a contour $\gamma : [a, b] \rightarrow \mathbb{C}$ is given by $\int_{\gamma} |dz|$.
- 11.11. Consider the contour integral $\int_{\Gamma} |z| dz$, where Γ is the upper half of the circle of radius r centered at the origin oriented clockwise. Compute the integral in two different ways:
- (i) Parametrize Γ and calculate the integral from the definition.
 - (ii) Note that $\int_{\Gamma} |z| dz = \int_{\Gamma} r dz$. Use the Cauchy–Goursat theorem to show that $\int_{\Gamma} r dz = \int_{-r}^r r dt$ and compute this last integral.
- 11.12. Evaluate the integral $\oint_C (z - \bar{z}) dz$ around the circle $C = \{z \in \mathbb{C} : |z| = 1\}$.
- 11.13. The function $f(z) = \frac{1}{z^2+1}$ fails to be holomorphic at $z = \pm i$. Let $\gamma : [0, 2\pi] \rightarrow \mathbb{C}$ be given by $\gamma(t) = re^{it}$ for some $r > 1$. Prove that $\oint_{\gamma} f(z) dz = 0$ by following these steps:
- (i) Use the Cauchy–Goursat theorem to show that the answer is independent of $r > 1$.
 - (ii) Bound the norm of $f(z)$ on γ in terms of r .
 - (iii) Use this bound to show that the absolute value of the integral is less than ε , if r is large enough. Beware: It generally does not make sense to talk about inequalities of the form $\int_{\gamma} |f| dz \leq \int_{\gamma} |g| dz$, since neither term is necessarily real.
-

- 11.14. Let γ be the curve $1 + 2e^{it}$ for $t \in [0, 2\pi]$. Compute the following:

- (i) $\int_{\gamma} \frac{(z+1)e^z}{z} dz$.
- (ii) $\int_{\gamma} \frac{(z+1)e^z}{z^3} dz$.
- (iii) $\int_{\gamma} \frac{\cos(z)}{z^2+2} dz$.
- (iv) $\int_{\gamma} \frac{\cos(z)}{(z^2+2)^2} dz$.

$$(v) \int_{\gamma} \frac{\sin(z)}{z^2 - z} dz.$$

$$(vi) \int_{\gamma} \frac{\sin(z)}{(z^2 - z)^2} dz.$$

- 11.15. Reevaluate the integral of Exercise 11.12 using the Cauchy integral formula. Hint: On C we have $1 = |z|^2 = z\bar{z}$, which implies that $\bar{z} = \frac{1}{z}$ on C .
- 11.16. Reevaluate the integral of Exercise 11.13 using the Cauchy integral formula and changing the contour to two small circles, one around i and one around $-i$.
- 11.17. The n th Legendre polynomial $P_n(z)$ is defined to be

$$P_n(z) = \frac{1}{2^n n!} \frac{d^n}{dz^n} [(z^2 - 1)^n].$$

Prove that

$$P_n(z) = \frac{1}{2\pi i} \oint_{\gamma} \frac{(w^2 - 1)^n}{2^n (w - z)^{n+1}} dw,$$

where γ is any simple closed curve containing z .

- 11.18. Let $f(z)$ be holomorphic such that $|f(z)| \leq M$ for all z with $|z - z_0| \leq r$. Prove that

$$|f^{(n)}(z_0)| \leq M n! r^{-n}.$$

- 11.19. Prove that any function f that is holomorphic on a punctured ball $B(z_0, r) \setminus \{z_0\}$ and is continuous at z_0 must, in fact, be holomorphic on the whole ball $B(z_0, r)$, following these steps:
- Choose a circle γ around z_0 of radius less than r , and define a function $F(z) = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(w)}{w - z} dw$. Prove that $F(z)$ is holomorphic inside of γ .
 - Prove that $f(z) = F(z)$ on the interior of γ , except possibly at z_0 .
 - Show that $f(z_0) = F(z_0)$, so $F = f$ on the interior of γ .

- 11.20. Liouville's theorem guarantees that all nonconstant holomorphic functions are unbounded on $\mathbb{C}$.
- Find a sequence $z_1, z_2, \dots$ such that $|\sin(z_n)| \rightarrow \infty$. Hint: $\sin(z) = \frac{1}{2}i(e^{-iz} - e^{iz})$.
 - Find a sequence $w_1, w_2, \dots$ such that $|\cos(w_n)| \rightarrow \infty$.
 - Consider the function $f(z) = \frac{1}{z^3 - 1}$. Prove that $|f| \rightarrow 0$ as $|z| \rightarrow \infty$. This function is not constant. Why isn't that a counterexample to Liouville's theorem?
- 11.21. A common form of the fundamental theorem of algebra states that any polynomial with coefficients in $\mathbb{C}$ of degree n has exactly n roots in $\mathbb{C}$, counted with multiplicity. Use Theorem 11.5.4 to prove this alternative form of the fundamental theorem of algebra.
- 11.22. The *minimum modulus principle* states that if f is holomorphic and not constant on a path-connected, open set U , and if $|f(z)| \neq 0$ for every $z \in U$, then $|f|$ has no minimum on U .
- Prove the minimum modulus principle.

- (ii) Give an example of a nonconstant, holomorphic function on an open set U such that $|f(z_0)| = 0$ for some $z_0 \in U$.
- (iii) Use your example from the previous step to show why the condition $|f| \neq 0$ is necessary in the proof.

11.23. Let $D = \{z \in \mathbb{C} : |z| \leq 1\}$ be the unit disk. Let $f: D \rightarrow D$ be holomorphic on the open unit ball $B(0, 1) \subset \mathbb{C}$, with $f(0) = 0$. In this exercise we prove that $|f'(0)| \leq 1$ and that $|f(z)| \leq |z|$ for all $z \in B(0, 1)$. We further show that if there is any nonzero $z_0 \in B(0, 1)$ such that $|f(z_0)| = |z_0|$, or if $|f'(0)| = 1$, then $f(z) = \alpha z$ for some α with $|\alpha| = 1$ (so f is a rotation).

We do this in several steps:

- (i) Let

$$g(z) = \begin{cases} f(z)/z & \text{if } z \neq 0, \\ f'(0) & \text{if } z = 0. \end{cases}$$

Prove that g is holomorphic away from 0 and continuous at 0.

- (ii) Use Exercise 11.19 to deduce that g is holomorphic on all of $B(0, 1)$.
- (iii) Show that for any $r < 1$ the maximum value of $|g|$ on $\overline{B(0, r)}$ must be attained on the circle $|z| = r$. Use this to show that $|f(z)| \leq |z|$ on $B(0, 1)$. Hint: Show $|g(z)| \leq 1/r$ for all $r < 1$.
- (iv) Show that if $|f(z_0)| = |z_0|$ anywhere in $B(0, 1) \setminus \{0\}$, then $f(z) = \alpha z$ for some $|\alpha| = 1$.
- (v) Show that $|f'(0)| \leq 1$ and if $|f'(0)| = 1$, then $f(z) = \alpha z$ with $|\alpha| = 1$.

11.24. For each of the following functions, find the Laurent expansion on an annulus of the form $\{z \mid 0 < |z - z_0| < R\}$ around the specified point z_0 , and find the greatest R for which the expansion converges

- (i) $z/(z - 1)$ around $z_0 = 0$.
- (ii) $e^z/(z - 1)$ around $z_0 = 1$.
- (iii) $\sin(z)/z$ around $z_0 = 0$.
- (iv) e^{z^3} around $z_0 = 0$.

11.25. Prove that if $f = \sum_{k=0}^{\infty} a_k(z - z_0)^k$ and $g = \sum_{j=0}^{\infty} b_j(z - z_0)^j$ are $\mathbb{C}$ -valued and both have radius of convergence at least r , then the product

$$\sum_{k=0}^{\infty} a_k(z - z_0)^k \sum_{j=0}^{\infty} b_j(z - z_0)^j = \sum_{n=0}^{\infty} \sum_{k+j=n} a_k b_j (z - z_0)^n$$

has radius of convergence at least r . Hint: Use Taylor's theorem and induction to show that the n th coefficient of the power series expansion of fg is $\sum_{k+j=n} a_k b_j$.

11.26. Find the Laurent expansion of the following functions in the indicated region:

- (i) $z/(z + 1)$ in the region $0 < |z| < 1$.
- (ii) e^z/z^2 in the region $0 < |z| < \infty$.
- (iii) $\frac{1-2z}{z(z-1)(z-3)^2}$ in the region $0 < |z - 3| < 1$.

11.27. Uniqueness of the Laurent expansion depends on the choice of region. That is, different regions for the Laurent expansion give different expansions. Show this by computing the Laurent expansion for $f(z) = \frac{1}{z(z-1)}$ in the following two regions:

(i) $0 < |z| < 1$.

(ii) $1 < |z| < \infty$.

11.28. Prove Proposition 11.6.9.

11.29. Prove Proposition 11.7.8.

11.30. Prove Lemma 11.7.12.

11.31. Compute the following residues:

(i)

$$\frac{1-2z}{z(z-1)(z-3)^2} \quad \text{at } z=0.$$

(ii)

$$\frac{1-2z}{z(z-1)(z-3)^2} \quad \text{at } z=1.$$

(iii)

$$\frac{1-2z}{z(z-1)(z-3)^2} \quad \text{at } z=3.$$

11.32. Find the following integrals:

(i)

$$\frac{1}{2\pi i} \oint_{|z|=\frac{1}{2}} \frac{1-2z}{z(z-1)(z-3)^2} dz.$$

(ii)

$$\frac{1}{2\pi i} \oint_{|z-1|=\frac{1}{2}} \frac{1-2z}{z(z-1)(z-3)^2} dz.$$

(iii)

$$\frac{1}{2\pi i} \oint_{|z-3|=\frac{1}{2}} \frac{1-2z}{z(z-1)(z-3)^2} dz.$$

(iv)

$$\frac{1}{2\pi i} \oint_{|z|=4} \frac{1-2z}{z(z-1)(z-3)^2} dz.$$

11.33. Find the real integral

$$\int_{-\infty}^{\infty} \frac{x^2}{1+x^4} dx = \lim_{R \rightarrow \infty} \int_{-R}^R \frac{x^2}{1+x^4} dx$$

by considering the contour integral $\oint_{\gamma} \frac{z^2}{1+z^4} dz$, where γ is the contour consisting of the union of the upper half of the circle of radius R and the real interval $[-R, R]$ (that is, $\gamma = \{z \in \mathbb{C} \mid R = |z|, \Im(z) \geq 0\} \cup [-R, R]$). Hint: Bound $\frac{z^2}{1+z^4}$ on the upper half circle and show that this part of the integral goes to 0 as $R \rightarrow \infty$.

11.34.* For each function $p(z)$, find the number of zeros inside of the region A without explicitly solving for them.

(i) $p(z) = z^6 + 4z^2 - 1$ and A is the disk $|z| < 1$.

(ii) $p(z) = z^3 + 9z + 27$ and A is the disk $|z| < 2$.

(iii) $p(z) = z^6 - 5z^2 + 10$ and A is the annulus $1 < |z| < 2$.

(iv) $p(z) = z^4 - z + 5$ and A is the first quadrant $\{z \in \mathbb{C} \mid \Re(z) > 0, \Im(z) > 0\}$.

11.35.* Show that $e^z = 5z^3 - 1$ has exactly three solutions in the unit ball $B(0, 1)$.

11.36.* Prove the following extension of the zero and pole counting formula.

Let f be a $\mathbb{C}$ -valued meromorphic function on a simply connected open set U with zeros $z_1, \dots, z_m$ of multiplicities $a_1, \dots, a_m$, respectively, and poles $w_1, \dots, w_n$ of multiplicities $b_1, \dots, b_n$, respectively. If none of the zeros or poles of f lie on a closed contour γ , and if h is holomorphic on U , then

$$\oint_{\gamma} \frac{f'(z)}{f(z)} h(z) dz = 2\pi i \left(\sum_{k=1}^m a_k h(z_k) I(\gamma, z_k) - \sum_{k=1}^n b_k h(w_k) I(\gamma, w_k) \right).$$

11.37.* Let γ be the circle $|z| = R$ of radius R (traversed once, counterclockwise).

Let $f(z) = z^n + a_{n-1}z^{n-1} + \dots + a_1z + a_0$, where the coefficients a_k are all complex numbers (and notice that the leading coefficient is 1). Show that

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi i} \oint_{\gamma} \frac{f'(z)}{f(z)} z dz = -a_{n-1}$$

in the following steps:

(i) If $\lambda_1, \dots, \lambda_n$ are all the zeros of f , show that the sum satisfies

$$\sum_{j=1}^n \lambda_j = -a_{n-1}.$$

(ii) Use Exercise 11.36 to show that

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi i} \oint_{\gamma} \frac{f'(z)}{f(z)} z dz = \sum_{j=1}^n \lambda_j.$$

11.38.* Use the holomorphic open mapping theorem to give a new proof of the maximum modulus principle.

Notes

Much of our treatment of complex analysis is inspired by the books [MH99] and [Der72]. Sections 11.6–11.8 are especially indebted to [MH99]. Other references include [Ahl78, CB09, SS02, Tay11]. Our proof of the Cauchy–Goursat theorem is modeled after [CB09, Sect. 47] and [Cos12]. Exercise 11.3 is from [SS02].

Part IV

Linear Analysis II

12 Spectral Calculus

Trying to make a model of an atom by studying its spectrum is like trying to make a model of a grand piano by listening to the noise it makes when thrown downstairs.
—British Journal of Radiology

In this chapter, we return to the study of spectral theory for linear operators on finite-dimensional vector spaces, or, equivalently, spectral theory of matrices. In Chapter 4 we proved several key results about spectral theory, but largely restricted our study to semisimple (and hence diagonalizable) matrices. The tools of complex analysis developed in the previous chapter give more powerful techniques for studying the spectral properties of matrices. In particular, we can now generalize the results of Chapter 4 to general matrix operators.

The main innovation in this chapter is the use of a matrix-valued function called the *resolvent* of an operator. Given $A \in M_n(\mathbb{F})$, the resolvent is the function $R(z) = (zI - A)^{-1}$. The first main result is the *spectral resolution formula*, which says that if f is any holomorphic function whose power series converges on a disk containing the entire spectrum of A , then the operator $f(A)$ can be computed using the Cauchy integral formula

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z)R(z)dz,$$

where Γ is a suitably chosen simple curve. As an immediate corollary of the spectral resolution formula, we get the famous *Cayley–Hamilton theorem*, which says that the characteristic polynomial $p(z)$ of A satisfies $p(A) = 0$.

The next main result is the *spectral decomposition* of an operator, which says that the space $\mathbb{F}^n$ can be decomposed as a direct sum of generalized eigenspaces of the operator, and that the operator can be decomposed in terms of how it acts on those eigenspaces. The spectral decomposition leads to easy proofs of several key results, including the *power method* for computing the dominant eigenvalue-eigenvector pair and the *spectral mapping theorem*, which says that when an operator A is mapped by a holomorphic function, the eigenvalues of $f(A)$ are just the images $f(\lambda)$, where $\lambda \in \sigma(A)$. It also gives a nice way to write out the spectral decomposition of $f(A)$ in terms of the spectral decomposition of A .

Finally, we conclude the chapter with some applications of the spectral decomposition and the spectral mapping theorem. These include *Perron's theorem*, the *Drazin inverse*, and the *Jordan normal form* of a matrix. Perron's theorem tells us about the existence and uniqueness of a dominant eigenvalue of a positive matrix operator. This theorem, combined with the power method, lies at the heart of some interesting modern applications, including Google's PageRank algorithm for ranking web page search results. The *Drazin inverse* is another pseudoinverse, different from the Moore–Penrose inverse. The Moore–Penrose inverse is a geometric inverse, determined in part by the inner product, whereas the Drazin inverse is a spectral inverse, determined by the eigenprojections of the particular operator.

The Jordan normal form of a matrix arises from a special choice of basis that allows the spectral decomposition to be written in a nearly diagonal form. This form has eigenvalues along the diagonal, zeros and ones on the superdiagonal, and zeros elsewhere. The Jordan form can sometimes be a useful approach when dealing with “fun-sized” matrices, that is, matrices small enough that you might decompose them by hand in an effort to gain intuition. But it is important to emphasize that the basis-independent form of the spectral decomposition, as described in Theorem 12.6.12, is almost always more useful than the Jordan normal form. Limiting yourself to a particular basis (in this case the Jordan basis) can be a hindrance to problem solving. Nevertheless, for those who really wish to use the Jordan normal form, our previous results about the spectral decomposition give an easy proof of the Jordan decomposition.

12.1 Projections

A *projection* is a special type of operator whose domain can be decomposed into two complementary subspaces: one which maps identically to itself, and one which maps to zero. A standard example of a projection is the operator $p : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ given by $P(x, y, z) = (x, y, 0)$. Intuitively, the image of an object in $\mathbb{R}^3$ under this projection is the shadow in the plane that the object would cast if the sun were directly overhead (on the z -axis), hence the name “projection.” We have already seen many examples of this type of projection in Section 3.1.3, namely, orthogonal projections. In this section, we consider more general projections that may or may not be orthogonal.⁴⁴

Throughout this section assume that V is a vector space.

12.1.1 Projections

Definition 12.1.1. A linear operator $P : V \rightarrow V$ is a projection if $P^2 = P$.

Example 12.1.2. It is straightforward to show that if P is a projection, then so is the *complementary projection* $I - P$; see Exercise 12.1.

⁴⁴Nonorthogonal projections are sometimes called *oblique* projections to distinguish them from orthogonal projections.

Lemma 12.1.3. *If P is a projection, then*

- (i) $\mathbf{y} \in \mathcal{R}(P)$ if and only if $P\mathbf{y} = \mathbf{y}$;
- (ii) $\mathcal{N}(P) = \mathcal{R}(I - P)$.

Proof.

- (i) If $\mathbf{y} \in \mathcal{R}(P)$, then $\mathbf{y} = P\mathbf{x}$ for some $\mathbf{x}$, which implies $P\mathbf{y} = P(P\mathbf{x}) = P\mathbf{x} = \mathbf{y}$. The converse is trivial.
- (ii) Note that $\mathbf{x} \in \mathcal{N}(P)$ if and only if $P\mathbf{x} = 0$, and this holds if and only if $(I - P)\mathbf{x} = \mathbf{x}$. This is equivalent to $\mathbf{x} \in \mathcal{R}(I - P)$, using (i). $\square$

A projection separates its domain into two complementary subspaces—the kernel and the range.

Theorem 12.1.4. *If P is a projection on V , then $V = \mathcal{R}(P) \oplus \mathcal{N}(P)$.*

Proof. For every $\mathbf{x} \in V$ we have $\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x}$, which gives $V = \mathcal{R}(P) + \mathcal{N}(P)$. It suffices to show that $\mathcal{R}(P) \cap \mathcal{N}(P) = \{0\}$. If $\mathbf{x} \in \mathcal{R}(P) \cap \mathcal{N}(P)$, then $P\mathbf{x} = \mathbf{x}$ and $P\mathbf{x} = 0$. Therefore $\mathbf{x} = 0$. $\square$

Corollary 12.1.5. *Assume V is finite dimensional. If P is a projection on V with $S = [\mathbf{s}_1, \dots, \mathbf{s}_k]$ as a basis for $\mathcal{R}(P)$ and $T = [\mathbf{t}_1, \dots, \mathbf{t}_\ell]$ as a basis for $\mathcal{N}(P)$, then $S \cup T$ is a basis for V , and the block matrix representation of P in the basis $S \cup T$ is*

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

where I is the $k \times k$ identity matrix (on $\mathcal{R}(P)$), and the zero blocks are submatrices of the appropriate sizes.

Proof. By Theorem 12.1.4 the basis $S \cup T$ is a basis of V . By Theorem 4.2.9 the matrix representation in this basis is of the form

$$\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} is the matrix representation of P on $\mathcal{R}(P)$ and A_{22} is the matrix representation of P on $\mathcal{N}(P)$. But P is zero on $\mathcal{N}(P)$, so A_{22} is the zero matrix. By Lemma 12.1.3(i) we know P is the identity on all vectors in $\mathcal{R}(P)$; hence, A_{11} is the identity matrix. $\square$

Theorem 12.1.6. *If W_1 and W_2 are complementary subspaces of V , that is, $V = W_1 \oplus W_2$, then there exists a unique projection P on V satisfying $\mathcal{R}(P) = W_1$ and $\mathcal{N}(P) = W_2$. In this case we say that P is the projection onto W_1 along W_2 .*

Proof. For each $\mathbf{x} \in V$ there exists a unique $\mathbf{x}_1 \in W_1$ and $\mathbf{x}_2 \in W_2$ such that $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$. Define the map P by setting $P\mathbf{x} = \mathbf{x}_1$ for each $\mathbf{x} \in V$. Since $P\mathbf{x}_1 = \mathbf{x}_1$ and $P\mathbf{x}_2 = \mathbf{0}$, we can see that $P^2 = P$ on both $\mathcal{R}(P)$ and $\mathcal{N}(P)$, separately, and hence on all of V . To show that P is linear, we let $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ and $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{y}_1 \in W_1$ and $\mathbf{x}_2, \mathbf{y}_2 \in W_2$. Since $a\mathbf{x} + b\mathbf{y} = a\mathbf{x}_1 + b\mathbf{y}_1 + a\mathbf{x}_2 + b\mathbf{y}_2$, we have $P(a\mathbf{x} + b\mathbf{y}) = a\mathbf{x}_1 + b\mathbf{y}_1 = aP(\mathbf{x}) + bP(\mathbf{y})$. To show uniqueness, suppose there exists some other projection Q satisfying $\mathcal{R}(Q) = W_1$ and $\mathcal{N}(Q) = W_2$. Since $P\mathbf{x} = P\mathbf{x}_1 + P\mathbf{x}_2 = \mathbf{x}_1 = Q\mathbf{x}_1 = Q\mathbf{x}_1 + Q\mathbf{x}_2 = Q\mathbf{x}$ for all $\mathbf{x}$, we have $Q = P$. $\square$

Example 12.1.7. Let $V = \mathbb{F}[x]$. If $W_1 = \mathbb{F}[x; n]$ is the space of polynomials of degree no more than n and W_2 is the space of polynomials having no monomials of degree n or less, then $\mathbb{F}[x] = W_1 \oplus W_2$. The projection of $\mathbb{F}[x]$ onto W_1 along W_2 is the map

$$\sum_{i=0}^m a_i x^i \mapsto \sum_{i=0}^n a_i x^i$$

that just forgets the terms of degree more than n .

12.1.2 Invariant Subspaces and Their Projections

In this section we introduce some results that combine the invariant subspaces of Section 4.2 with the results of the previous section on projections. In addition to being important for our treatment of spectral theory, these results are useful in numerical linear algebra, differential equations, and control theory.

Theorem 12.1.8. *Let $L : V \rightarrow V$ be a linear operator. A subspace W of V is L -invariant if and only if for any projection P onto W we have $LP = PLP$.*

Proof. ($\implies$) Assume W is L -invariant and $\mathcal{R}(P) = W$. Since $\mathcal{N}(I - P) = W$, it follows that $(I - P)L\mathbf{w} = \mathbf{0}$ for all $\mathbf{w} \in W$, and also $P\mathbf{v} \in W$ for all $\mathbf{v} \in V$. Therefore, we have that $(I - P)LP\mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in V$ and $LP = PLP$.

($\impliedby$) Assume that $LP = PLP$ and that $\mathcal{R}(P) = W$. Since $P\mathbf{w} = \mathbf{w}$ for all $\mathbf{w} \in W$, it follows that $L\mathbf{w} = PL\mathbf{w}$ for all $\mathbf{w} \in W$. Thus, $L\mathbf{w} \in W$ for all $\mathbf{w} \in W$. Therefore, W is L -invariant. $\square$

Theorem 12.1.9. *Assume that L is a linear operator on V . Two complementary subspaces W_1 and W_2 of V satisfy $LW_1 \subset W_1$ and $LW_2 \subset W_2$ (we call these mutually L -invariant) if and only if the corresponding projection P onto W_1 along W_2 satisfies $LP = PL$.*

Proof. ($\implies$) Since W_1 and W_2 are complementary, any $\mathbf{v} \in V$ can be uniquely written as $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$. Since W_1 and W_2 are mutually L -invariant and P satisfies $\mathcal{R}(P) = W_1$ and $\mathcal{N}(P) = W_2$, we have

$$PL\mathbf{v} = PL\mathbf{w}_1 + PL\mathbf{w}_2 = PL\mathbf{w}_1 = L\mathbf{w}_1 = LP\mathbf{w}_1 = LP\mathbf{v}.$$

($\Leftarrow$) Assume that $PL = LP$, where P is a projection satisfying $\mathcal{R}(P) = W_1$ and $\mathcal{N}(P) = W_2$. If $\mathbf{w}_1 \in W_1$, then $L\mathbf{w}_1 = LP\mathbf{w}_1 = PL\mathbf{w}_1 \in W_1$. If $\mathbf{w}_2 \in W_2$, then $L\mathbf{w}_2 = L(I - P)\mathbf{w}_2 = (I - P)L\mathbf{w}_2 \in W_2$. Thus, W_1 and W_2 are mutually L -invariant. $\square$

12.1.3 Eigenprojections for Simple Operators

Some of the most important projections are those with range equal to the eigenspace of an operator. In this section we consider such projections for simple matrices (see Definition 4.3.4). It is straightforward to generalize these results to semisimple matrices, but, in fact, they can also be generalized to matrices that are not semisimple. Much of the rest of this chapter is devoted to developing this generalization.

Proposition 12.1.10. *Let $A \in M_n(\mathbb{F})$ have distinct eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. Let $\mathbf{r}_1, \dots, \mathbf{r}_n$ be a corresponding basis of (right) eigenvectors, and let S be the matrix whose columns are these eigenvectors, so the i th column of S is $\mathbf{r}_i$. Let $\ell_1^\top, \dots, \ell_n^\top$ be the rows of S^{-1} (the left eigenvectors of A , by Remark 4.3.19), and define the rank-1 map $P_i = \mathbf{r}_i \ell_i^\top$. For all i, j we have*

- (i) $\ell_i^\top \mathbf{r}_j = \delta_{ij}$,
- (ii) $P_i P_j = \delta_{ij} P_i$,
- (iii) $P_i A = A P_i = \lambda_i P_i$,
- (iv) $\sum_{i=1}^n P_i = I$, and
- (v) $A = \sum_{i=1}^n \lambda_i P_i$.

We call the maps P_i the eigenprojections of the simple matrix A .

Proof.

- (i) This follows from the fact that $S^{-1}S = I$.
- (ii) $P_i P_j = \mathbf{r}_i \ell_i^\top \mathbf{r}_j \ell_j^\top = \delta_{ij} \mathbf{r}_i \ell_j^\top = \delta_{ij} P_i$.
- (iii) $P_i A = \mathbf{r}_i \ell_i^\top A = \lambda_i \mathbf{r}_i \ell_i^\top = \lambda_i P_i$ and $A P_i = A \mathbf{r}_i \ell_i^\top = \lambda_i \mathbf{r}_i \ell_i^\top = \lambda_i P_i$.
- (iv) $\sum_{i=1}^n P_i = \sum_{i=1}^n \mathbf{r}_i \ell_i^\top$, which is the outer-product expansion of $SS^{-1} = I$.
- (v) This follows by combining (iii) and (iv). $\square$

The proposition says that any simple operator L on an n -dimensional space can be decomposed into the sum of n rank-1 eigenprojections, and we can decompose its domain into the direct sum of one-dimensional invariant subspaces. Representing L

in terms of the eigenbasis $\mathbf{r}_1, \dots, \mathbf{r}_n$ gives a diagonal matrix. In this representation the eigenvectors are just the standard basis vectors $\mathbf{r}_i = \mathbf{e}_i$, and the projections P_i are

$$P_i = \mathbf{e}_i \mathbf{e}_i^\top = \begin{bmatrix} 0 & 0 & & \cdots & 0 \\ 0 & 0 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & 1 & \cdots & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & & & & & 0 \end{bmatrix},$$

where the only nonzero entry is $p_{ii} = 1$. In this setting it is clear that the matrix representation of L is

$$A = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \lambda_1 P_1 + \cdots + \lambda_n P_n.$$

The previous proposition says how to express these projections in terms of any basis—not just an eigenbasis. In other words, the eigenprojections give a “basis-free” description of spectral theory for simple operators (or matrices). In the rest of the chapter we will develop a similar basis-free version of spectral theory for matrices that are not necessarily even semisimple.

Example 12.1.11. Let

$$A = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}.$$

We saw in Example 4.1.12 that the eigenvalues are $\lambda_1 = -2$ and $\lambda_2 = 5$, with corresponding eigenvectors

$$\mathbf{r}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{r}_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}.$$

It is not hard to check that

$$\ell_1^\top = \frac{1}{7} \begin{bmatrix} -4 & 3 \end{bmatrix} \quad \text{and} \quad \ell_2^\top = \frac{1}{7} \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Thus the eigenprojections are

$$P_1 = \frac{1}{7} \begin{bmatrix} 4 & -3 \\ -4 & 3 \end{bmatrix} \quad \text{and} \quad P_2 = \frac{1}{7} \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix}.$$

These are both rank 1, and we can check the various properties in the lemma:

(ii) We have

$$P_1 P_2 = \frac{1}{49} \begin{bmatrix} 4 & -3 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and it is similarly straightforward to check that $P_1^2 = P_1$ and $P_2^2 = P_2$.

(iii) We have

$$P_1 A = \frac{1}{7} \begin{bmatrix} 4 & -3 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} -8 & 6 \\ 8 & -6 \end{bmatrix} = \lambda_1 P_1,$$

and a similar computation shows that $P_2 A = \lambda_2 P_2$.

(iv) We have

$$P_1 + P_2 = \frac{1}{7} \begin{bmatrix} 4 & -3 \\ -4 & 3 \end{bmatrix} + \frac{1}{7} \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I.$$

(v) Finally,

$$\lambda_1 P_1 + \lambda_2 P_2 = \frac{-2}{7} \begin{bmatrix} 4 & -3 \\ -4 & 3 \end{bmatrix} + \frac{5}{7} \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} = A.$$

12.2 Generalized Eigenvectors

Recall from Theorem 4.3.7 that a matrix $A \in M_n(\mathbb{F})$ has a basis of eigenvectors if and only if it is diagonalizable. This allows us to write the domain of A as a direct sum of eigenspaces. But the eigenvectors of a matrix that is not semisimple do not span the domain of the matrix, so if we want to decompose such an operator like we did for simple operators, we must generalize the idea of an eigenspace to include more vectors.

In this section, we do just that. We develop a generalization of an eigenspace that allows us to write the domain of any matrix operator as a direct sum of *generalized eigenspaces*. Later in the chapter we use this to transform any matrix operator into a nice block-diagonal form.

12.2.1 The Index of an Operator

In order to discuss generalized eigenspaces, we need the concept of the *index* of an operator. Recall from Exercise 2.8 that for any linear operator B , we have an increasing sequence of subspaces

$$\mathcal{N}(B) \subset \mathcal{N}(B^2) \subset \cdots \subset \mathcal{N}(B^k) \subset \cdots \quad (12.1)$$

If the operator is defined on a finite-dimensional space, then this sequence must eventually stabilize. In other words, there exists a $K \in \mathbb{N}$ such that $\mathcal{N}(B^k) = \mathcal{N}(B^{k+1})$ for all $k \geq K$.

Definition 12.2.1. The index of a matrix $B \in M_n(\mathbb{F})$ is the smallest $k \in \mathbb{N}$ such that $\mathcal{N}(B^k) = \mathcal{N}(B^{k+1})$. We denote the index of B by $\text{ind}(B)$.

Example 12.2.2. For any operator B we always have $B^0 = I$, so if B is invertible, then $\text{ind}(B) = 0$ because the kernels of $B^0 = I$ and $B^1 = B$ are both trivial: $\mathcal{N}(B^0) = \mathcal{N}(B^1) = \{0\}$.

In the ascending chain (12.1) it might seem possible for there to be some proper (strict) inclusions that could be followed by an equality and then some more proper inclusions before the terminal subspace. The following theorem shows that this is not possible—once we have an equality, the rest are all equalities.

Theorem 12.2.3. The index of any $B \in M_n(\mathbb{F})$ is well defined and finite. Furthermore, if $k = \text{ind}(B)$, then $\mathcal{N}(B^m) = \mathcal{N}(B^k)$ for all $m \geq k$, and every inclusion $\mathcal{N}(B^{l-1}) \subset \mathcal{N}(B^l)$ for $l = 1, 2, \dots, k$ is proper.

Proof. Since B is an operator on a finite-dimensional space, only a finite number of inclusions in (12.1) can be proper, so the index is well defined and finite. The rest of the theorem follows immediately from Exercise 2.12. $\square$

Example 12.2.4. Consider the matrix

$$B = \begin{bmatrix} 0 & -1 & 1 & 1 \\ 3 & -2 & -4 & 5 \\ 1 & -1 & -1 & 2 \\ 2 & -1 & -3 & 3 \end{bmatrix}.$$

It is straightforward to check that $B^2 = 0$, so $B^k = 0$ for all $k \geq 2$. But because $B \neq 0$ we have $\dim \mathcal{N}(B) < \dim \mathcal{N}(B^2)$. Hence, $\text{ind}(B) = 2$.

The rank-nullity theorem says that the dimension of the range and the kernel of any operator on $\mathbb{F}^n$ must add to n , but it is possible that these spaces may have a nontrivial intersection. The next theorem shows that when we take k large enough, the range and kernel of B^k have a trivial intersection, and so together they span the entire space.

Theorem 12.2.5. Let $B \in M_n(\mathbb{F})$. If $k \geq \text{ind}(B)$, then $\mathcal{N}(B^k) \cap \mathcal{R}(B^k) = \{0\}$ and $\mathbb{F}^n = \mathcal{N}(B^k) \oplus \mathcal{R}(B^k)$.

Proof. Suppose $\mathbf{x} \in \mathcal{N}(B^k) \cap \mathcal{R}(B^k)$. Thus, $B^k \mathbf{x} = \mathbf{0}$, and there exists $\mathbf{y} \in \mathbb{F}^n$ such that $\mathbf{x} = B^k \mathbf{y}$. Hence, $B^k B^k \mathbf{y} = B^{2k} \mathbf{y} = \mathbf{0}$, and so $\mathbf{y} \in \mathcal{N}(B^{2k}) = \mathcal{N}(B^k)$. Therefore $\mathbf{x} = B^k \mathbf{y} = \mathbf{0}$. It follows that $\mathcal{N}(B^k) \cap \mathcal{R}(B^k) = \{\mathbf{0}\}$.

To prove that $\mathbb{F}^n = \mathcal{N}(B^k) + \mathcal{R}(B^k)$, use the rank-nullity theorem, which states that $n = \dim \mathcal{N}(B^k) + \text{rank}(B^k)$. By Corollary 2.3.19, we have $\mathbb{F}^n = \mathcal{N}(B^k) + \mathcal{R}(B^k)$. Therefore, $\mathbb{F}^n = \mathcal{N}(B^k) \oplus \mathcal{R}(B^k)$. $\square$

Corollary 12.2.6. *If $B \in M_n(\mathbb{F})$ and $k = \text{ind}(B)$, then $\mathcal{R}(B^m) = \mathcal{R}(B^k)$ for all $m \geq k$.*

Proof. By Exercise 2.8 we also have $\mathcal{R}(B^m) \subset \mathcal{R}(B^k)$, so it suffices to show $\dim B^m = \dim B^k$. But by Theorem 12.2.3 we have $\mathcal{N}(B^m) = \mathcal{N}(B^k)$; so, Theorem 12.2.5 implies that $\dim B^m = \dim B^k$. $\square$

We conclude this section with one final important observation about repeated powers of an operator B on a vector $\mathbf{x}$.

Proposition 12.2.7. *Assume $B \in M_n(\mathbb{F})$. If $B^m \mathbf{x} = \mathbf{0}$ and $B^{m-1} \mathbf{x} \neq \mathbf{0}$ for some $m \in \mathbb{Z}^+$, then the set $\{\mathbf{x}, B\mathbf{x}, \dots, B^{m-1}\mathbf{x}\}$ is linearly independent.*

Proof. Suppose that $a_0 \mathbf{x} + a_1 B\mathbf{x} + \dots + a_{m-1} B^{m-1} \mathbf{x} = \mathbf{0}$ is a nontrivial linear combination. Let a_i be the first nonzero coefficient. Left multiplying by B^{m-i-1} gives $a_i B^{m-1} \mathbf{x} = \mathbf{0}$. But $B^{m-1} \mathbf{x} \neq \mathbf{0}$ implies $a_i = 0$, a contradiction. $\square$

12.2.2 Generalized Eigenspaces

Recall that an eigenspace of an operator L associated with the eigenvalue λ is the subspace $\Sigma_\lambda = \mathcal{N}(\lambda I - L)$, where the geometric multiplicity of λ is the dimension of the eigenspace. If $A \in M_n(\mathbb{F})$ and A is semisimple, then we can write $\mathbb{F}^n$ as the direct sum of A -invariant subspaces

$$\mathbb{F}^n = \mathcal{N}(\lambda_1 I - A) \oplus \mathcal{N}(\lambda_2 I - A) \oplus \dots \oplus \mathcal{N}(\lambda_r I - A),$$

where $\sigma(A) = \{\lambda_i\}_{i=1}^r$ is the set of distinct eigenvalues of A . Choosing a basis for each $\mathcal{N}(\lambda_i I - A)$ and writing the operator corresponding to A in terms of the resulting basis for $\mathbb{F}^n$ yields a diagonal matrix that is similar to A , where each block is of the form $\lambda_i I$.

If A is not diagonalizable, the spaces $\mathcal{N}(\lambda_i I - A)$ do not span $\mathbb{F}^n$, but as explained in the previous section, we have a nested sequence of subspaces

$$\mathcal{N}(\lambda I - A) \subset \mathcal{N}((\lambda I - A)^2) \subset \dots \subset \mathcal{N}((\lambda I - A)^n) \subset \dots \quad (12.2)$$

If $k = \text{ind}(\lambda I - A)$, then the sequence stabilizes at $\mathcal{N}((\lambda I - A)^k)$. For each eigenvalue λ , the space $\mathcal{N}((\lambda I - A)^k)$ is a good generalization of the eigenspace. This generalization allows us to put a nondiagonalizable operator L into block-diagonal form. This is an important approach for understanding the spectral theory of finite-dimensional operators.

Definition 12.2.8. If $A \in M_n(\mathbb{F})$ with eigenvalue $\lambda \in \mathbb{C}$, then the generalized eigenspace of A corresponding to λ is the subspace $\mathcal{E}_\lambda = \mathcal{N}((\lambda I - A)^k)$, where $k = \text{ind}(\lambda I - A)$. Any nonzero element of $\mathcal{E}_\lambda$ is called a generalized eigenvector of A corresponding to λ .

The next four lemmata provide the key tools we need to prove the main theorem of this section (Theorem 12.2.14), which says $\mathbb{F}^n$ decomposes into a direct sum of its generalized eigenspaces.

Lemma 12.2.9. If $A \in M_n(\mathbb{F})$ and λ is an eigenvalue of A , then the generalized eigenspace $\mathcal{E}_\lambda$ is A -invariant.

Proof. Any subspace is invariant under the operator λI . Since $A = \lambda I - (\lambda I - A)$, it suffices to show that $\mathcal{E}_\lambda$ is $(\lambda I - A)$ -invariant. If $\mathbf{x} \in \mathcal{E}_\lambda$ and $\mathbf{y} = (\lambda I - A)\mathbf{x}$, then $(\lambda I - A)^k \mathbf{y} = (\lambda I - A)^{k+1} \mathbf{x} = \mathbf{0}$, which implies that $\mathbf{y} \in \mathcal{E}_\lambda$. $\square$

Example 12.2.10. Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{bmatrix}.$$

This matrix has two eigenvalues, $\lambda = 1$ and $\lambda = 3$, but is not semisimple. A straightforward calculation shows that the eigenvectors associated with $\lambda = 1$ and $\lambda = 3$ are $[1 \ 0 \ 0]^T$ and $[1 \ 2 \ 2]^T$, respectively. An additional calculation shows that $\text{ind}(3I - A) = 1$ and $\text{ind}(1I - A) = 2$. So to span the generalized eigenspace $\mathcal{E}_1$, we need an eigenvector corresponding to $\lambda = 1$ and one additional generalized eigenvector $\mathbf{v}_3 \in \mathcal{N}((1I - A)^2)$ with $\mathbf{v}_3 \notin \mathcal{N}(1I - A)$. The vector $\mathbf{v}_3 = [0 \ 1 \ 0]^T$ satisfies this condition. Thus the generalized eigenspaces are $\mathcal{E}_1 = \text{span}(\{[1 \ 2 \ 2]^T, [0 \ 1 \ 0]^T\})$ and $\mathcal{E}_3 = \text{span}(\{[1 \ 0 \ 0]^T\})$.

Lemma 12.2.11. If λ and μ are distinct eigenvalues of a matrix $A \in M_n(\mathbb{F})$, then $\mathcal{E}_\lambda \cap \mathcal{E}_\mu = \{\mathbf{0}\}$.

Proof. Let $k_\lambda = \text{ind}(\lambda I - A)$ and $k_\mu = \text{ind}(\mu I - A)$. The previous lemma shows that $\mathcal{E}_\lambda$ is $(\lambda I - A)$ -invariant. We claim that $\mu I - A$ restricted to $\mathcal{E}_\lambda$ has a trivial kernel, and thus (by iterating) the kernel of $(\mu I - A)^{k_\mu}$ on $\mathcal{E}_\lambda$ is also trivial. This implies that $\mathcal{E}_\lambda \cap \mathcal{E}_\mu = \{\mathbf{0}\}$.

To prove the claim, suppose that $\mathbf{x} \in \mathcal{N}(\mu I - A) \cap \mathcal{E}_\lambda$, so that $A\mathbf{x} = \mu\mathbf{x}$ and $(\lambda I - A)^{k_\lambda} \mathbf{x} = \mathbf{0}$. Using the binomial theorem to expand $(\lambda I - A)^{k_\lambda}$ shows that $(\lambda - \mu)^{k_\lambda} \mathbf{x} = \mathbf{0}$; see Exercise 12.9 for details. This implies that $\mathbf{x} = \mathbf{0}$. $\square$

Lemma 12.2.12. *Assume that W_1 and W_2 are A -invariant subspaces of $\mathbb{F}^n$ with $W_1 \cap W_2 = \{\mathbf{0}\}$. If λ is an eigenvalue of A with generalized eigenspace $\mathcal{E}_\lambda$, such that $\mathcal{E}_\lambda \cap W_i = \{\mathbf{0}\}$ for each i , then*

$$\mathcal{E}_\lambda \cap (W_1 \oplus W_2) = \{\mathbf{0}\}.$$

Proof. If $\mathbf{x} \in \mathcal{E}_\lambda \cap (W_1 \oplus W_2)$, then $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, with $\mathbf{x}_i \in W_i$ for each i . If $k = \text{ind}(\lambda I - A)$, then

$$\mathbf{0} = (\lambda I - A)^k \mathbf{x} = (\lambda I - A)^k \mathbf{x}_1 + (\lambda I - A)^k \mathbf{x}_2.$$

Since each W_i is A -invariant, it is also $(\lambda I - A)$ -invariant, so we have

$$(\lambda I - A)^k \mathbf{x}_1 = -(\lambda I - A)^k \mathbf{x}_2 \in W_1 \cap W_2 = \{\mathbf{0}\}.$$

Therefore, for each i we have $\mathbf{x}_i \in \mathcal{N}((\lambda I - A)^k) = \mathcal{E}_\lambda$. But we also have $\mathbf{x}_i \in W_i$ by definition, and thus $\mathbf{x}_i = \mathbf{0}$. This implies that $\mathbf{x} = \mathbf{0}$. $\square$

Lemma 12.2.13. *If $A \in M_n(\mathbb{F})$ and λ is an eigenvalue of A , then $\dim(\mathcal{E}_\lambda)$ equals the algebraic multiplicity m_λ of λ .*

Proof. The proof follows that of Theorem 4.4.5. By Schur's lemma we can assume that A is upper triangular and of the form

$$\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

where the block T_{11} is upper triangular with all diagonal values equal to λ , and the block T_{22} is upper triangular with all diagonal values being different than λ . The block $\lambda I - T_{11}$ is strictly upper triangular, and the block $\lambda I - T_{22}$ is upper triangular but has all nonzero diagonals and is thus nonsingular. It follows that $(\lambda I - T_{11})^{m_\lambda} = 0$ and $(\lambda I - T_{22})^k$ is nonsingular for all $k \in \mathbb{N}$. Therefore, $\dim(\mathcal{E}_\lambda) = \dim \mathcal{N}((\lambda I - A)^{m_\lambda}) = \dim \mathcal{N}((\lambda I - T_{11})^{m_\lambda}) = m_\lambda$. $\square$

Theorem 12.2.14. *Given $A \in M_n(\mathbb{F})$, we can decompose $\mathbb{F}^n$ into a direct sum of A -invariant generalized eigenspaces*

$$\mathbb{F}^n = \mathcal{E}_{\lambda_1} \oplus \mathcal{E}_{\lambda_2} \oplus \cdots \oplus \mathcal{E}_{\lambda_r}, \quad (12.3)$$

where $\sigma(A) = \{\lambda_1, \lambda_2, \dots, \lambda_r\}$ are the distinct eigenvalues of A .

Proof. We claim first that for any $\lambda \in \sigma(A)$ and for any subset $M = \{\mu_1, \dots, \mu_\ell\} \subset \sigma(A)$ of distinct eigenvalues with $\lambda \notin M$, we have

$$\mathcal{E}_\lambda \cap \bigoplus_{\mu \in M} \mathcal{E}_\mu = \{\mathbf{0}\}.$$

To show this we use Lemmata 12.2.11, 12.2.9, and 12.2.12 and induct on the size of the set M .

If $|M| = 1$, the claim holds by Lemma 12.2.11. If the claim holds for $|M| = m$, then for any M' with $|M'| = m + 1$, write $M' = M \cup \{\nu\}$. Set $W_0 = \mathcal{E}_\lambda$, set $W_1 = \mathcal{E}_\nu$, and set $W_2 = \bigoplus_{\mu \in M} \mathcal{E}_\mu$. The conditions of Lemma 12.2.12 hold, and so $\mathcal{E}_\lambda \cap \bigoplus_{\mu \in M'} \mathcal{E}_\mu = \{0\}$, proving the claim.

This shows we can define the subspace $W = \mathcal{E}_{\lambda_1} \oplus \mathcal{E}_{\lambda_2} \oplus \cdots \oplus \mathcal{E}_{\lambda_r}$ of $\mathbb{F}^n$. So, it suffices to show that $W = \mathbb{F}^n$. By Lemma 12.2.13, the dimensions of the generalized eigenspaces are equal to the algebraic multiplicities, which add up to n . This implies that $\dim W = n$, which implies $W = \mathbb{F}^n$ by Corollary 1.4.7. $\square$

Remark 12.2.15. By Theorem 12.2.14, a given matrix A with eigenvalues $\sigma(A) = \{\lambda_1, \lambda_2, \dots, \lambda_r\}$ is similar to a block-diagonal matrix where the generalized eigenspace $\mathcal{E}_{\lambda_i}$ is invariant with respect to the corresponding block. In Section 12.10, we describe a method for producing a specific basis for each generalized eigenspace so that the block-diagonal matrix is banded with the eigenvalues on the diagonal, zeros and ones on the superdiagonal, and zeros elsewhere. This is called the *Jordan canonical form*. The rest of this chapter develops a basis-free approach for producing the block-diagonal representation, where we use projections instead of making an explicit choice of generalized eigenvectors.

12.3 The Resolvent

In this section we introduce the *resolvent*, which is a powerful tool for understanding the spectral properties of an operator. The resolvent allows us to use the tools of complex analysis, including the Cauchy integral formula, to study properties of an operator.

One of the many things the resolvent gives us is a complete set of generalized eigenprojections that describes the spectral theory of the operator in a basis-free way. The resolvent is also a key tool in the proof of the power method (see Section 12.7.3) and the Perron–Frobenius theorem (Theorem 12.8.11), which are essential for working with Markov chains, and which are fundamental to applications like the Google PageRank algorithm (see Section 12.8.3).

12.3.1 Properties of the Resolvent

Definition 12.3.1. Let $A \in M_n(\mathbb{F})$. The resolvent set $\rho(A) \subset \mathbb{C}$ of A consists of the points $z \in \mathbb{C}$ for which $(zI - A)^{-1}$ exists. The complement $\sigma(A) = \mathbb{C} \setminus \rho(A)$ is called the spectrum of A . The resolvent⁴⁵ of A is the map $R(A, \cdot) : \rho(A) \rightarrow M_n(\mathbb{C})$, given by

$$R(A, z) = (zI - A)^{-1}. \quad (12.4)$$

If there is no ambiguity, we denote $R(A, z)$ simply as $R(z)$.

⁴⁵Although we define and study the resolvent only for matrices (i.e., finite-dimensional operators), many of the results in this chapter work in the much more general case of *closed* linear operators (see, for example, [Sch12].)

Example 12.3.2. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}. \quad (12.5)$$

By direct computation we get the resolvent

$$R(z) = \begin{bmatrix} z-1 & -2 \\ -3 & z \end{bmatrix}^{-1} = \frac{1}{(z-3)(z+2)} \begin{bmatrix} z & 2 \\ 3 & z-1 \end{bmatrix}. \quad (12.6)$$

Note that $R(z)$ has poles precisely at $\sigma(A) = \{3, -2\}$, so the resolvent set is $\rho(A) = \mathbb{C} \setminus \{3, -2\}$.

Remark 12.3.3. Cramer's rule (Corollary 2.9.24) shows that the resolvent takes the form of the rational function

$$R(z) = (zI - A)^{-1} = \frac{\text{adj}(zI - A)}{\det(zI - A)}, \quad (12.7)$$

where $\text{adj}(\cdot)$ is the adjugate matrix. Since the denominator of (12.7) is the characteristic polynomial of A , the poles of $R(z)$ correspond precisely, in terms of both location and multiplicity, to the eigenvalues of A . In other words, the spectrum is the set of eigenvalues of A , and the resolvent set $\rho(A)$ is precisely those points of $\mathbb{C}$ that are not eigenvalues of A .

Example 12.3.4. Let

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 7 \end{bmatrix}.$$

Using (12.7) we compute the resolvent as $R(z) = B(z)/p(z)$, where

$$B(z) = \begin{bmatrix} (z-1)^2(z-7) & 3(z-1)(z-7) & 9(z-7) & 27 \\ 0 & (z-1)^2(z-7) & 3(z-1)(z-7) & 9(z-1) \\ 0 & 0 & (z-1)^2(z-7) & 3(z-1)^2 \\ 0 & 0 & 0 & (z-1)^3 \end{bmatrix}$$

and $p(z) = (z-1)^3(z-7)$.

Lemma 12.3.5. Let $A \in M_n(\mathbb{F})$. The following identities hold:

(i) If $z_1, z_2 \in \rho(A)$, then

$$R(z_2) - R(z_1) = (z_1 - z_2)R(z_2)R(z_1). \quad (12.8)$$

This is called Hilbert's identity.

(ii) If $z \in \rho(A_1) \cap \rho(A_2)$, then

$$R(A_2, z) - R(A_1, z) = R(A_1, z)(A_2 - A_1)R(A_2, z). \quad (12.9)$$

(iii) If $z \in \rho(A)$, then

$$R(z)A = AR(z). \quad (12.10)$$

(iv) If $z_1, z_2 \in \rho(A)$, then

$$R(z_1)R(z_2) = R(z_2)R(z_1). \quad (12.11)$$

Proof.

- (i) Write $(z_1 - z_2)I = (z_1I - A) - (z_2I - A)$, and then left and right multiply by $R(z_2)$ and $R(z_1)$, respectively.
- (ii) Write $A_2 - A_1 = (zI - A_1) - (zI - A_2)$, and then left and right multiply by $R(A_1, z)$ and $R(A_2, z)$, respectively.
- (iii) From (12.4), we have $(zI - A)R(z) = R(z)(zI - A)$, which simplifies to (12.10).
- (iv) For $z_1 = z_2$, (12.11) follows trivially. Otherwise, using (12.8) and relabeling indices, we have

$$R(z_2)R(z_1) = \frac{R(z_2) - R(z_1)}{z_1 - z_2} = \frac{R(z_1) - R(z_2)}{z_2 - z_1} = R(z_1)R(z_2). \quad \square$$

12.3.2 Local Properties

Throughout the remainder of this section, let $A \in M_n(\mathbb{F})$, and let $\|\cdot\|$ be a matrix⁴⁶ norm.

Theorem 12.3.6. *The set $\rho(A)$ is open, and $R(z)$ is holomorphic on $\rho(A)$ with the following convergent power series at $z_0 \in \rho(A)$ for $|z - z_0| < \|R(z_0)\|^{-1}$:*

$$R(z) = \sum_{k=0}^{\infty} (-1)^k (z - z_0)^k R^{k+1}(z_0). \quad (12.12)$$

Proof. We use the fact (from Proposition 5.7.4) that when $\|B\| < 1$ the Neumann series $\sum_{k=0}^{\infty} B^k$ converges to $(I - B)^{-1}$. From (12.8) we have

$$R(z_0) = R(z) + (z - z_0)R(z_0)R(z) = [I + (z - z_0)R(z_0)]R(z).$$

Setting $B = -(z - z_0)R(z_0)$ in the Neumann series, we have

$$R(z) = [I + (z - z_0)R(z_0)]^{-1}R(z_0) = \sum_{k=0}^{\infty} (-1)^k (z - z_0)^k R^{k+1}(z_0),$$

and this series converges in the open neighborhood $\{z \in \mathbb{C} \mid |z - z_0| < \|R(z_0)\|^{-1}\}$ of z_0 . Therefore $\rho(A)$ is an open set, and $R(z)$ is holomorphic on $\rho(A)$. $\square$

⁴⁶Recall from Definition 3.5.15 that a matrix norm is a norm on $M_n(\mathbb{F})$ that satisfies the submultiplicative property $\|AB\| \leq \|A\|\|B\|$.

Remark 12.3.7. Comparing (12.12) with the Taylor series (11.18) for $R(z)$ reveals a relationship between powers of $R(z)$ and its derivatives:

$$\frac{d^k}{dz^k} R(z) = k!(-1)^k R^{k+1}(z) = k!(-1)^k (zI - A)^{-(k+1)}. \quad (12.13)$$

We can also get this by formally taking derivatives of $R(z)$.

Theorem 12.3.8. *If $|z| > \|A\|$, then $R(z)$ exists and is given by*

$$R(z) = \sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}}. \quad (12.14)$$

Proof. Note that

$$R(z) = (zI - A)^{-1} = z^{-1}(I - z^{-1}A)^{-1} = z^{-1} \sum_{k=0}^{\infty} \frac{A^k}{z^k},$$

which converges whenever $\|z^{-1}A\| < 1$. $\square$

Corollary 12.3.9.

$$\lim_{|z| \rightarrow \infty} \|R(z)\| = 0. \quad (12.15)$$

Moreover, $R(z)$ is holomorphic in a neighborhood of $z = \infty$.

Remark 12.3.10. When we say $R(z)$ is holomorphic in a neighborhood of ∞ , we simply mean that if we make the substitution $w = 1/z$, then $R(1/w)$ is holomorphic in a neighborhood of $w = 0$.

Proof. That $R(z)$ is holomorphic in a neighborhood of $z = \infty$ follows immediately from (12.14). For $|z| > \|A\|$ we have

$$\|R(z)\| \leq \sum_{k=0}^{\infty} \frac{\|A\|^k}{|z|^{k+1}} = \frac{1}{|z|} \left(1 - \frac{\|A\|}{|z|}\right)^{-1} = \frac{1}{|z| - \|A\|}.$$

This shows $\lim_{|z| \rightarrow \infty} \|R(z)\| = 0$. $\square$

Remark 12.3.11. Remark 12.3.3 shows that the spectrum is the set of roots of the characteristic polynomial. By the fundamental theorem of algebra (Theorem 11.5.4), this is nonempty and finite; see also Exercise 11.21.

The previous theorem gives an alternative way to see that the spectrum is nonempty, which we give in the following corollary.

Corollary 12.3.12. *The spectrum $\sigma(A)$ is nonempty.*

Proof. Suppose that $\sigma(A) = \emptyset$, so that $\rho(A) = \mathbb{C}$ and $R(z)$ is entire. By Corollary 12.3.9, $R(z)$ is uniformly bounded on $\mathbb{C}$. Hence, $R(z)$ is constant by Liouville's theorem (Theorem 11.5.1). Thus, $R(z) = \lim_{z \rightarrow \infty} R(z) = 0$. This is a contradiction, since $I = (zI - A)R(z)$. $\square$

Lemma 12.3.13. *For any matrix norm $\|\cdot\|$ and any $A \in M_n(\mathbb{F})$, the limit*

$$r(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k} \quad (12.16)$$

exists and $r(A) \leq \|A\|$. We call $r(A)$ the spectral radius of A .

Proof. For $1 \leq m < k$ fixed, we have the inequality $\|A^k\| \leq \|A^{k-m}\| \|A^m\|$. Let $a_k = \log \|A^k\|$. The inequality implies that $a_k \leq a_m + a_{k-m}$. By the division algorithm, there exists a unique q and r such that $k = qm + r$, where $0 \leq r < m$, or, alternatively, we have $q = \lfloor k/m \rfloor$. This gives

$$a_k = a_{qm+r} \leq qa_m + a_r.$$

It follows that

$$\frac{a_k}{k} \leq \frac{q}{k} a_m + \frac{1}{k} a_r.$$

Leaving m fixed and letting $k > m$ grow gives

$$\limsup_k \frac{a_k}{k} = \limsup_k \frac{\lfloor k/m \rfloor}{k} a_m = \frac{a_m}{m}.$$

It follows that

$$\limsup_k \frac{a_k}{k} \leq \limsup_k \left(\frac{q}{k} a_m + \frac{1}{k} a_r \right) = \frac{a_m}{m}. \quad (12.17)$$

Since (12.17) holds for all m , it follows that

$$\limsup_k \frac{a_k}{k} \leq \liminf_m \frac{a_m}{m}.$$

Therefore the limit exists. To show that $r(A) \leq \|A\|$, let $m = 1$ in (12.17). $\square$

Theorem 12.3.14. *The regions of convergence in Theorems 12.3.6 and 12.3.8 can be increased to*

- (i) $|z - z_0| < [r(R(z_0))]^{-1}$ and
- (ii) $|z| > r(A)$, respectively.

Proof. In each part below, let r denote $r(R(z_0))$ and $r(A)$, respectively.

- (i) Let z satisfy $|z - z_0| < r^{-1}$. There exists an $\varepsilon > 0$ such that $|z - z_0| < (r + 2\varepsilon)^{-1}$. Moreover, there exists N such that $\|R(z_0)^k\|^{1/k} < r + \varepsilon$ for all $k > N$. Thus $|z - z_0|^k \|R(z_0)^k\| \leq \left(\frac{r + \varepsilon}{r + 2\varepsilon} \right)^k < 1$, which implies that (12.12) converges.

- (ii) Let z satisfy $|z| > r$. There exists an $\varepsilon > 0$ such that $|z| > r + 2\varepsilon$. Moreover there exists N such that $\|A^k\|^{1/k} < r + \varepsilon$ for all $k > N$. Thus $|z|^{-k} \|A^k\| < (\frac{r+\varepsilon}{r+2\varepsilon}) < 1$, which implies that (12.14) converges. $\square$

Remark 12.3.15. Theorem 12.3.14 implies that $R(z)$ is holomorphic on $|z| > r(A)$, which implies

$$r(A) \geq \sigma_M = \sup_{\lambda \in \sigma(A)} |\lambda|.$$

In the next section we show in the finite-dimensional case that $r(A) = \sigma_M$. Not only does this justify the name spectral radius for $r(A)$, but it also implies that the value of $r(A)$ is independent of the operator norm used in (12.16).

12.4 Spectral Resolution

Throughout this section assume $A \in M_n(\mathbb{F})$ is given. Recall from (12.7) that $R(z)$ is a matrix of rational functions with a common denominator. The poles of these functions correspond, in terms of both location and algebraic multiplicity, to the eigenvalues of A . In Section 12.3.2, we saw the locally holomorphic behavior of $R(z)$ in $\rho(A)$. In the next three sections we determine the Laurent expansion of $R(z)$ about the points of $\sigma(A)$.

Definition 12.4.1. Let $\lambda \in \sigma(A)$. Let Γ be a positively oriented simple closed curve containing $\lambda \in \sigma(A)$ but no other points of $\sigma(A)$. The spectral projection (or eigenprojection) of A associated with λ is given by

$$P_\lambda = \text{Res}(R(z), \lambda) = \frac{1}{2\pi i} \oint_\Gamma R(z) dz. \quad (12.18)$$

Theorem 12.4.2. For any $A \in M_n(\mathbb{C})$ and any $\lambda \in \sigma(A)$, the following properties of P_λ hold:

- (i) *Idempotence:* $P_\lambda^2 = P_\lambda$ for all $\lambda \in \sigma(A)$.
- (ii) *Independence:* $P_\lambda P_{\lambda'} = P_{\lambda'} P_\lambda = 0$, whenever $\lambda, \lambda' \in \sigma(A)$ with $\lambda \neq \lambda'$.
- (iii) *A-invariance:* $AP_\lambda = P_\lambda A$ for all $\lambda \in \sigma(A)$.
- (iv) *Completeness:* $\sum_{\lambda \in \sigma(A)} P_\lambda = I$.

Proof.

- (i) Let Γ and Γ' be two positively oriented simple closed contours in $\rho(A)$ surrounding λ and no other points of $\sigma(A)$. Assume also that Γ is interior to Γ' as in Figure 12.1. For each $z \in \Gamma$ we have

$$\oint_{\Gamma'} \frac{dz'}{z' - z} = 2\pi i \quad \text{and} \quad \oint_\Gamma \frac{dz}{z' - z} = 0.$$

By Hilbert's identity we have

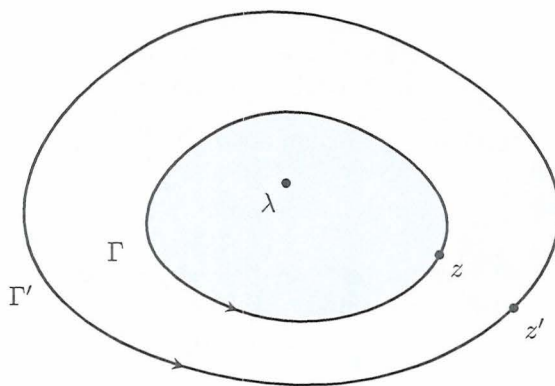


Figure 12.1. Illustration to assist in the proofs of Theorem 12.4.2(i) and Lemmata 12.5.2 and 12.5.3. The point z is interior to Γ' , but z' is exterior to Γ .

$$\begin{aligned}
 P_\lambda^2 &= \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma} \oint_{\Gamma'} R(z) R(z') dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma} \oint_{\Gamma'} \frac{R(z) - R(z')}{z' - z} dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \left[\oint_{\Gamma} R(z) \left(\oint_{\Gamma'} \frac{dz'}{z' - z} \right) dz - \oint_{\Gamma'} R(z') \left(\oint_{\Gamma} \frac{dz}{z' - z} \right) dz' \right] \\
 &= \frac{1}{2\pi i} \oint_{\Gamma} R(z) dz = P_\lambda.
 \end{aligned}$$

- (ii) Let Γ and Γ' be two positively oriented disjoint simple closed contours in $\rho(A)$ surrounding λ and λ' , respectively, and such that no other points of $\sigma(A)$ are interior to either curve, as in Figure 12.2. Note that

$$\oint_{\Gamma'} \frac{dz'}{z' - z} = 0 \quad \text{and} \quad \oint_{\Gamma} \frac{dz}{z' - z} = 0.$$

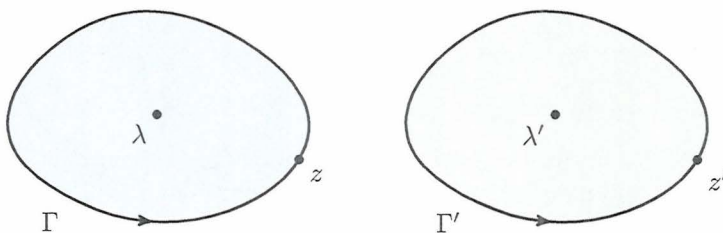


Figure 12.2. Diagram to assist in the proof of Theorem 12.4.2(ii). In this case z is not interior to Γ' , nor is z' interior to Γ , and no points of $\sigma(A) \setminus \{\lambda, \lambda'\}$ are interior to either curve.

Again using Hilbert's identity, we have

$$\begin{aligned}
 P_\lambda P_{\lambda'} &= \left(\frac{1}{2\pi i} \right)^2 \oint_\Gamma \oint_{\Gamma'} R(z) R(z') dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \oint_\Gamma \oint_{\Gamma'} \frac{R(z) - R(z')}{z' - z} dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \left[\oint_\Gamma R(z) \left(\oint_{\Gamma'} \frac{dz'}{z' - z} \right) dz - \oint_{\Gamma'} R(z') \left(\oint_\Gamma \frac{dz}{z' - z} \right) dz' \right] \\
 &= 0.
 \end{aligned}$$

(iii) This follows directly from (12.10), that is,

$$A \left[\oint_\Gamma R(z) dz \right] = \oint_\Gamma A R(z) dz = \oint_\Gamma R(z) A dz = \left[\oint_\Gamma R(z) dz \right] A.$$

(iv) Let Γ be a positively oriented circle centered at $z = 0$ having radius $R > r(A)$. Theorems 12.3.8 and 12.3.14 guarantee that the Laurent expansion $R(z) = \sum_{k=0}^{\infty} A^k z^{-(k+1)}$ holds along Γ , so the residue theorem (Theorem 11.7.13) applied to this Laurent series gives

$$\frac{1}{2\pi i} \oint_\Gamma R(z) dz = \frac{1}{2\pi i} \oint_\Gamma \sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}} dz = A^0 = I.$$

However, the residue theorem applied to $R(z)$ gives

$$\frac{1}{2\pi i} \oint_\Gamma R(z) dz = \sum_{\lambda \in \sigma(A)} \frac{1}{2\pi i} \oint_{\Gamma_\lambda} R(z) dz = \sum_{\lambda \in \sigma(A)} P_\lambda,$$

where each Γ_λ is a positively oriented simple closed contour containing λ and no other points of $\sigma(A)$. $\square$

Remark 12.4.3. The double integrals in the previous proof require Fubini's theorem (Theorem 8.6.1) to change the order of integration. Although we only proved Fubini's theorem for real-valued integrals, it is straightforward to extend it to the complex-valued case.

Example 12.4.4. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}$$

of Example 12.3.2. The partial fraction decomposition of the resolvent (12.6) is

$$R(z) = \frac{1}{z-3} \left(\frac{1}{5} \begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} \right) + \frac{1}{z+2} \left(\frac{1}{5} \begin{bmatrix} 2 & -2 \\ -3 & 3 \end{bmatrix} \right).$$

Taking residues gives the spectral projections

$$P_3 = \frac{1}{5} \begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} \quad \text{and} \quad P_{-2} = \frac{1}{5} \begin{bmatrix} 2 & -2 \\ -3 & 3 \end{bmatrix}.$$

Notice that all four parts of Theorem 12.4.2 hold.

Example 12.4.5. Let

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

be the matrix in Example 12.3.4. Use (12.18) to compute P_λ for $\lambda = 7$, and note that $P_1 = I - P_7$. We get

$$P_7 = \begin{bmatrix} 0 & 0 & 0 & 1/8 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad P_1 = \begin{bmatrix} 1 & 0 & 0 & -1/8 \\ 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Theorem 12.4.6 (Spectral Resolution Formula). Suppose that $f(z)$ has a power series at $z = 0$ with radius of convergence $b > r(A)$. For any positively oriented simple closed contour Γ containing $\sigma(A)$, we have

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z) R(z) dz. \quad (12.19)$$

Proof. Express $f(z)$ as a power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$. Without loss of generality, let Γ be a positively oriented circle centered at $z = 0$ having radius b_0 , where $r(A) < b_0 < b$. Thus, we have

$$\frac{1}{2\pi i} \oint_{\Gamma} f(z) R(z) dz = \frac{1}{2\pi i} \oint_{\Gamma} f(z) z^{-1} \sum_{k=0}^{\infty} \frac{A^k}{z^k} dz.$$

Since $f(z)z^{-1}$ is bounded on Γ and the summation converges uniformly on compact subsets, the sum and integral can be interchanged to give

$$\sum_{k=0}^{\infty} A^k \left[\frac{1}{2\pi i} \oint_{\Gamma} z^{-k-1} f(z) dz \right] = \sum_{k=0}^{\infty} a_k A^k = f(A). \quad \square$$

Corollary 12.4.7. The spectral radius of A is

$$r(A) = \sigma_M = \sup_{\lambda \in \sigma(A)} |\lambda|.$$

Proof. By Theorem 12.3.14, we know that $r(A) \geq \sigma_M$. For equality, it suffices to show that $r(A) \leq \sigma_M + \varepsilon$ for all $\varepsilon > 0$. Let Γ be a positively oriented circle centered at $z = 0$ of radius $\sigma_M + \varepsilon$. By the spectral resolution formula, we have

$$A^n = \frac{1}{2\pi i} \oint_{\Gamma} z^n R(z) dz. \quad (12.20)$$

Hence,

$$\|A^n\| \leq \frac{1}{2\pi} (\sigma_M + \varepsilon)^n K \cdot 2\pi(\sigma_M + \varepsilon) = K(\sigma_M + \varepsilon)^{n+1},$$

where

$$K = \sup_{\Gamma} \|R(z)\|_{\infty}.$$

This gives

$$r(A) = \lim_{n \rightarrow \infty} \|A^n\|^{1/n} \leq \lim_{n \rightarrow \infty} K^{1/n} (\sigma_M + \varepsilon)^{1+1/n} = \sigma_M + \varepsilon. \quad \square$$

Corollary 12.4.8 (Cayley–Hamilton Theorem). *If A is an $n \times n$ matrix with characteristic polynomial $p(z) = \det(zI - A)$, then $p(A) = 0$.*

Proof. Let Γ be a simple closed contour containing $\sigma(A)$. By the spectral resolution formula (12.19) and Cramer's rule (Corollary 2.9.24), we have

$$p(A) = \frac{1}{2\pi i} \oint_{\Gamma} \det(zI - A)(zI - A)^{-1} dz = \frac{1}{2\pi i} \oint_{\Gamma} \text{adj}(zI - A) dz = 0,$$

since $\text{adj}(zI - A)$ is a polynomial in z and hence holomorphic. $\square$

Nota Bene 12.4.9. It might seem tempting to try to prove the Cayley–Hamilton theorem by simply substituting A in for z in the expression $\det(zI - A)$. Unfortunately it doesn't even make sense to substitute a matrix A for the scalar z in the scalar multiplication zI . Contrast this with substituting A for z into $p(z)$, which is a polynomial in one variable, so $p(A)$ is a well-defined matrix in $M_n(\mathbb{C})$.

Example 12.4.10. Consider again the matrix A in Example 12.3.2. Note that the characteristic polynomial is $p(z) = z^2 - z - 6$. The Cayley–Hamilton theorem says that $p(A) = A^2 - A - 6I = 0$, which we verify directly:

$$\begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}^2 - \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix} - \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 7 & 2 \\ 3 & 6 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix} - \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

12.5 Spectral Decomposition I

In the previous section, we were able to use the residues of $R(z)$ at $\lambda \in \sigma(A)$ to obtain the spectral projections of A and prove their key properties. In this section and the next, we explore the rest of the Laurent expansion (11.20) of $R(z)$ at λ as a series of matrix operators, given by

$$R(z) = \sum_{k=-\infty}^{\infty} A_k(z - \lambda)^k. \quad (12.21)$$

According to Cauchy's integral theorem, each matrix operator A_k is given by (11.22)

$$A_k = \frac{1}{2\pi i} \oint_{\Gamma} \frac{R(z)}{(z - \lambda)^{k+1}} dz, \quad (12.22)$$

where Γ is a positively oriented simple closed contour containing λ and no other points of $\sigma(A)$; see Section 11.6.3 for a review of Laurent expansions.

The main goal of these two sections is to establish the *spectral decomposition formula* for any operator $A \in M_n(\mathbb{F})$. The spectral decomposition formula is the generalization of the formula $A = \sum_{\lambda} \lambda P_{\lambda}$ for semisimple matrices (see Proposition 12.1.10(v)) to general, not necessarily semisimple, matrices. We then show in Section 12.7 that the spectral decomposition is unique. Writing out the spectral decomposition explicitly in terms of a specific choice of basis gives the popular *Jordan normal form* of the operator, but our development is a basis-free description that works in any basis.

Nota Bene 12.5.1. Do not confuse the coefficient A_k of the Laurent expansion (12.21) of the resolvent of A with the power A^k of A . Both can be computed by a contour integral: A_k can be computed by (12.22) and A^k can be computed by (12.20) (an application of the spectral resolution formula). But despite the apparent similarity, they are very different things.

Lemma 12.5.2. Assume that $\lambda \in \sigma(A)$. Let Γ and Γ' be two positively oriented simple closed contours in $\rho(A)$ surrounding λ and no other points of $\sigma(A)$. Assume also that Γ is interior to Γ' , that z' is a point on Γ' , and that z is a point on Γ , as depicted in Figure 12.1. Let

$$\eta_n = \begin{cases} 1, & n \geq 0, \\ 0, & n < 0. \end{cases}$$

The following equalities hold:

$$(i) \quad \frac{1}{2\pi i} \oint_{\Gamma} (z - \lambda)^{-m-1} (z' - z)^{-1} dz = \eta_m (z' - \lambda)^{-m-1}, \quad (12.23)$$

$$(ii) \quad \frac{1}{2\pi i} \oint_{\Gamma'} (z' - \lambda)^{-n-1} (z' - z)^{-1} dz' = (1 - \eta_n) (z' - \lambda)^{-n-1}. \quad (12.24)$$

Proof.

- (i) Since z' is outside of Γ , the function $(z' - z)^{-1}$ is holomorphic within. Expand $(z' - z)^{-1}$ in terms of $z - \lambda$ to get

$$\frac{1}{z' - z} = \frac{1}{z' - \lambda} \cdot \frac{1}{1 - \left(\frac{z - \lambda}{z' - \lambda}\right)} = \sum_{k=0}^{\infty} \frac{(z - \lambda)^k}{(z' - \lambda)^{k+1}}.$$

Inserting into (12.23) and shrinking Γ to a small circle Γ_λ around λ with every point of Γ_λ nearer to λ than z' (see Figure 12.3) gives

$$\begin{aligned} & \frac{1}{2\pi i} \oint_{\Gamma_\lambda} (z - \lambda)^{-m-1} \left[\sum_{k=0}^{\infty} \frac{(z - \lambda)^k}{(z' - \lambda)^{k+1}} \right] dz \\ &= \sum_{k=0}^{\infty} (z' - \lambda)^{-k-1} \left[\frac{1}{2\pi i} \oint_{\Gamma_\lambda} (z - \lambda)^{-m-1+k} dz \right] \\ &= \eta_m (z' - \lambda)^{-m-1}. \end{aligned}$$

- (ii) In this case, both λ and z lie inside Γ . Split the contour into Γ_λ and Γ_z as in Figure 12.3, so the left side of (12.24) becomes

$$\begin{aligned} & \frac{1}{2\pi i} \oint_{\Gamma_\lambda} (z' - \lambda)^{-n-1} (z' - z)^{-1} dz' + \frac{1}{2\pi i} \oint_{\Gamma_z} (z' - \lambda)^{-n-1} (z' - z)^{-1} dz' \\ &= (1 - \eta_n) (z - \lambda)^{-n-1}. \end{aligned}$$

The first integral follows the same idea as (i), except with a minus sign. The second integral is an application of Cauchy's integral formula (11.14). $\square$

Lemma 12.5.3. *The coefficients of the Laurent expansion of $R(z)$ at $\lambda \in \sigma(A)$ satisfy the identity*

$$A_m A_n = (1 - \eta_m - \eta_n) A_{m+n+1}. \quad (12.25)$$

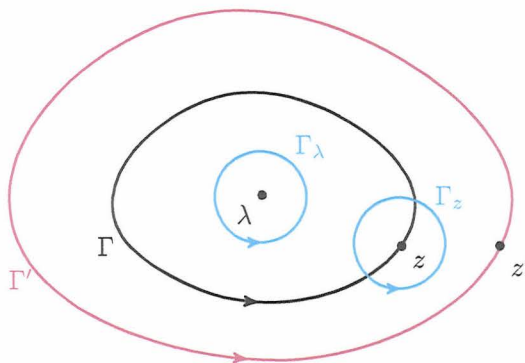


Figure 12.3. *Diagram of the paths and points in Lemma 12.5.2. The contour integral over Γ' (red) is the same as the sum of the integrals over the blue circles Γ_z and Γ_λ .*

Proof. Let Γ and Γ' be positively oriented simple closed contours surrounding λ and no other points of $\sigma(A)$. Assume also that Γ is interior to Γ' as depicted in Figure 12.1. We have

$$\begin{aligned}
 A_m A_n &= \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma} \oint_{\Gamma'} (z - \lambda)^{-m-1} (z' - \lambda)^{-n-1} R(z') R(z) dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma} \oint_{\Gamma'} (z - \lambda)^{-m-1} (z' - \lambda)^{-n-1} \frac{R(z) - R(z')}{z' - z} dz' dz \\
 &= \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma} (z - \lambda)^{-m-1} R(z) \left[\oint_{\Gamma'} (z' - \lambda)^{-n-1} (z' - z)^{-1} dz' \right] dz \\
 &\quad - \left(\frac{1}{2\pi i} \right)^2 \oint_{\Gamma'} (z' - \lambda)^{-n-1} R(z') \left[\oint_{\Gamma} (z - \lambda)^{-m-1} (z' - z)^{-1} dz \right] dz' \\
 &= \frac{1}{2\pi i} \oint_{\Gamma} (z - \lambda)^{-m-1} R(z) (1 - \eta_n) (z - \lambda)^{-n-1} dz \\
 &\quad - \frac{1}{2\pi i} \oint_{\Gamma'} (z' - \lambda)^{-n-1} R(z') \eta_m (z' - \lambda)^{-m-1} dz' \\
 &= (1 - \eta_m - \eta_n) \frac{1}{2\pi i} \oint_{\Gamma} (z - \lambda)^{-m-n-2} R(z) dz \\
 &= (1 - \eta_m - \eta_n) A_{m+n+1}. \quad \square
 \end{aligned}$$

Remark 12.5.4. Note that $P_{\lambda} = A_{-1}$, so Lemma 12.5.3 gives another proof that P_{λ} is a projection, since

$$P_{\lambda}^2 = A_{-1}^2 = A_{-1} = P_{\lambda}.$$

Lemma 12.5.5. Fix $\lambda \in \sigma(A)$ and define $D_{\lambda} = A_{-2}$ and $S_{\lambda} = A_0$. The following hold:

- (i) For $n \geq 2$, we have $A_{-n} = D_{\lambda}^{n-1}$.
- (ii) For $n \geq 1$, we have $A_n = (-1)^n S_{\lambda}^{n+1}$.
- (iii) The operator P_{λ} commutes with D_{λ} and S_{λ} :

$$P_{\lambda} D_{\lambda} = D_{\lambda} P_{\lambda} = D_{\lambda} \quad \text{and} \quad P_{\lambda} S_{\lambda} = S_{\lambda} P_{\lambda} = 0. \quad (12.26)$$

- (iv) The Laurent expansion of $R(z)$ around λ is

$$R(z) = \frac{P_{\lambda}}{z - \lambda} + \sum_{k=1}^{\infty} \frac{D_{\lambda}^k}{(z - \lambda)^{k+1}} + \sum_{k=0}^{\infty} (-1)^k (z - \lambda)^k S_{\lambda}^{k+1}. \quad (12.27)$$

- (v) We have

$$P_{\lambda} R(z) = R(z) P_{\lambda} = \frac{P_{\lambda}}{z - \lambda} + \sum_{k=1}^{\infty} \frac{D_{\lambda}^k}{(z - \lambda)^{k+1}}. \quad (12.28)$$

Proof. These follow by applying (12.25); see Exercises 12.23–25. $\square$

Example 12.5.6. Consider the matrix A given in Example 12.3.4. We found the eigenprojections in Example 12.4.5. Below we determine D_1 and S_1 . We then verify (12.27). Using (12.22) term by term for $\lambda = 1$ gives

$$A_{-2} = D_1 = 3 \begin{bmatrix} 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_{-3} = D_1^2 = 9 \begin{bmatrix} 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Moreover, $D_1^k = 0$ for $k > 2$. To find S_1 apply (12.27) and (12.28) to get

$$\sum_{k=0}^{\infty} (-1)^k (z-1)^k S_1^{k+1} = (I - P_1)R(z) = \frac{P_7}{z-7}.$$

Integrating shows that $S_1 = \frac{-1}{6}P_7$. Thus, for $k \in \mathbb{N}$ we have

$$S_1^{k+1} = \left(\frac{-1}{6}\right)^{k+1} P_7.$$

Putting this all together gives

$$R(z) = \frac{D_1^2}{(z-1)^3} + \frac{D_1}{(z-1)^2} + \frac{P_1}{z-1} + S_1 \sum_{k=0}^{\infty} \frac{(z-1)^k}{6^k}.$$

Note that the holomorphic part of the Laurent expansion is a geometric series, which sums nicely to give the final expression

$$R(z) = \frac{D_1^2}{(z-1)^3} + \frac{D_1}{(z-1)^2} + \frac{P_1}{z-1} + \frac{P_7}{z-7}.$$

12.6 Spectral Decomposition II

Throughout this section, let $A \in M_n(\mathbb{F})$ be a given matrix operator. The spectral resolution formula (Theorem 12.4.6) allows us to write $f(A)$ as a contour integral. In this section, we use that formula to prove the spectral decomposition theorem, which breaks up the contour integral into residue components P_λ and D_λ for each $\lambda \in \sigma(A)$. We also examine the properties of D_λ more carefully and show that D_λ is nilpotent. Hereafter, we refer to D_λ as the *eigennilpotent* of A associated with the eigenvalue λ .

Lemma 12.6.1. *For each $\lambda \in \sigma(A)$, the operator D_λ satisfies*

$$D_\lambda = (A - \lambda I)P_\lambda. \quad (12.29)$$

Moreover, its spectral radius $r(D_\lambda)$ is zero.

Proof. To prove (12.29) it suffices to prove that

$$AP_\lambda = \lambda P_\lambda + D_\lambda. \quad (12.30)$$

Let Γ_λ be a positively oriented circle around λ containing no other points of $\sigma(A)$. By definition, we have $R(z)(zI - A) = I$, so $zR(z) = AR(z) + I$. This gives

$$\begin{aligned} AP_\lambda &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} AR(z) dz \\ &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} AR(z) + Idz \quad (I \text{ is holomorphic}) \\ &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} zR(z) dz \\ &= \frac{1}{2\pi i} \oint_{\Gamma_\lambda} \lambda R(z) dz + \frac{1}{2\pi i} \oint_{\Gamma_\lambda} (z - \lambda)R(z) dz \\ &= \lambda P_\lambda + D_\lambda. \end{aligned}$$

To prove $r(D_\lambda) = 0$ parametrize Γ_λ by $z(t) = \lambda + \rho e^{it}$ for any sufficiently small choice of $\rho > 0$. By Lemma 12.5.5(i) we have

$$\begin{aligned} \|D_\lambda^k\| &= \left\| \frac{1}{2\pi i} \oint_{\Gamma_\lambda} (z - \lambda)^k R(z) dz \right\| \\ &= \frac{1}{2\pi} \left\| \int_0^{2\pi} \rho^k e^{ikt} R(z(t)) \rho i e^{it} dt \right\| \\ &\leq \rho^{k+1} \sup_{z \in \Gamma_\lambda} \|R(z)\|. \end{aligned}$$

Since Γ_λ is compact, set $M = \sup_{z \in \Gamma_\lambda} \|R(z)\|$, which is finite. Thus,

$$r(D_\lambda) = \lim_{k \rightarrow \infty} \|D_\lambda^k\|^{1/k} \leq \lim_{k \rightarrow \infty} \rho^{1+1/k} M^{1/k} = \rho.$$

Since we can choose ρ to be arbitrarily small, we must have $r(D_\lambda) = 0$. $\square$

Example 12.6.2. Let A be the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

of Examples 12.3.4 and 12.4.5. From Lemma 12.6.1 we compute

$$D_1 = (A - I)P_1 = \begin{bmatrix} 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -1/8 \\ 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix} = 3 \begin{bmatrix} 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

which agrees with our earlier computation in Example 12.5.6 and clearly has all its eigenvalues equal to 0. Moreover

$$D_7 = (A - 7I)P_7 = \begin{bmatrix} -6 & 3 & 0 & 0 \\ 0 & -6 & 3 & 0 \\ 0 & 0 & -6 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1/8 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Lemma 12.6.3. *A matrix $B \in M_n(\mathbb{F})$ satisfies $r(B) = 0$ if and only if B is nilpotent.*

Proof. If $r(A) = 0$, then $\sigma(A) = \{0\}$. Hence the characteristic polynomial of A is $p(\lambda) = \lambda^n$. By the Cayley–Hamilton theorem (Corollary 12.4.8), we have that $p(A) = A^n = 0$, which implies that A is nilpotent. Conversely, if A is nilpotent, then $A^k = 0$ for all $k \geq n$, which implies that $r(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k} = 0$. $\square$

Remark 12.6.4. Recall that the *order* of a nilpotent operator $B \in M_n(\mathbb{F})$ is the smallest m such that $B^m = 0$. Since the order m of a nilpotent operator is the same as its index, we have from Exercise 12.6 that $m \leq n$.

Proposition 12.6.5. *For each $\lambda \in \sigma(A)$, the order m_λ of the nilpotent operator D_λ satisfies $m_\lambda \leq \dim \mathcal{R}(P_\lambda)$.*

Proof. By Lemmata 12.6.1 and 12.6.3, the operator D_λ is nilpotent because $r(D_\lambda) = 0$. Since, $D_\lambda = P_\lambda D_\lambda = D_\lambda P_\lambda$, we can consider D_λ as an operator on $\mathcal{R}(P_\lambda)$. Thus by Remark 12.6.4, the order m_λ of D_λ satisfies $m_\lambda \leq \dim \mathcal{R}(P_\lambda)$. $\square$

Remark 12.6.6. The proposition implies that $R(z)$ is meromorphic, that is, it has no essential singularities. More precisely, (12.27) becomes

$$R(z) = \frac{P_\lambda}{z - \lambda} + \sum_{k=1}^{m_\lambda-1} \frac{D_\lambda^k}{(z - \lambda)^{k+1}} + \sum_{k=0}^{\infty} (-1)^k (z - \lambda)^k S_\lambda^{k+1}. \quad (12.31)$$

Therefore the principal part (12.28) becomes

$$P_\lambda R(z) = R(z) P_\lambda = \frac{P_\lambda}{z - \lambda} + \sum_{k=1}^{m_\lambda-1} \frac{D_\lambda^k}{(z - \lambda)^{k+1}}. \quad (12.32)$$

Lemma 12.6.7. *Let $\lambda \in \sigma(A)$ and $\mathbf{y} \in V$. If $(\lambda I - A)\mathbf{y} \in \mathcal{R}(P_\lambda)$, then $\mathbf{y} \in \mathcal{R}(P_\lambda)$.*

Proof. Assume $\mathbf{y} \neq 0$; otherwise the result is trivial. Let $\mathbf{v} = (\lambda I - A)\mathbf{y}$. If $\mathbf{v} \in \mathcal{R}(P_\lambda)$, then $\mathbf{v} = P_\lambda \mathbf{v}$. Independence of the projections (Theorem 12.4.2(ii)) implies that $P_\mu \mathbf{v} = 0$ whenever $\mu \in \sigma(A)$ and $\mu \neq \lambda$. Combining this with the fact that $P_\mu A = \mu P_\mu + D_\mu$ (12.30) and the fact that P_μ and D_μ commute (12.26) gives

$$0 = P_\mu (\lambda I - A)\mathbf{y} = \lambda P_\mu \mathbf{y} - \mu P_\mu \mathbf{y} - D_\mu \mathbf{y},$$

which implies

$$D_\mu P_\mu \mathbf{y} = D_\mu \mathbf{y} = (\lambda - \mu) P_\mu \mathbf{y}.$$

Since D_μ is nilpotent, it follows that $r(D_\mu) = 0$, which implies $\lambda = \mu$ (which is false) or $P_\mu \mathbf{y} = \mathbf{0}$. The fact that $I = \sum_{\mu \in \sigma(A)} P_\mu$ (see Theorem 12.4.2(iv)) gives

$$\mathbf{y} = \sum_{\mu \in \sigma(A)} P_\mu \mathbf{y} = P_\lambda \mathbf{y},$$

which implies $\mathbf{y} \in \mathcal{R}(P_\lambda)$ $\square$

Remark 12.6.8. The proof of the previous lemma did not use the definition of the projections P_μ and the nilpotents D_μ , but rather only the fact that the P_μ are projections satisfying the basic properties listed in Theorem 12.4.2, and that the D_μ are nilpotents satisfying commutativity of P_μ with D_μ and satisfying $P_\mu A = \mu P_\mu + D_\mu$. Thus, the lemma holds for any collection of projections and nilpotents indexed by the elements of $\sigma(A)$ and satisfying these properties.

Theorem 12.6.9. For each $\lambda \in \sigma(A)$, the generalized eigenspace $\mathcal{E}_\lambda$ is equal to $\mathcal{R}(P_\lambda)$.

Proof. We first show that $\mathcal{E}_\lambda \subset \mathcal{R}(P_\lambda)$. Recall that $\mathcal{E}_\lambda \subset \mathcal{N}((\lambda I - A)^{k_\lambda})$, where $k_\lambda = \text{ind}(\lambda I - A)$. Choose $\mathbf{x} \in \mathcal{E}_\lambda$ so that $(\lambda I - A)^{k_\lambda - 1} \mathbf{x} \neq \mathbf{0}$. The set $\{\mathbf{x}, (\lambda I - A)\mathbf{x}, \dots, (\lambda I - A)^{k_\lambda - 1} \mathbf{x}\}$ is a basis for $\mathcal{E}_\lambda$ by Proposition 12.2.7. It suffices to show that each basis vector is in $\mathcal{R}(P_\lambda)$.

If $\mathbf{y} = (\lambda I - A)^{k_\lambda - 1} \mathbf{x}$, then $(\lambda I - A)\mathbf{y} = \mathbf{0} \in \mathcal{R}(P_\lambda)$, which from Lemma 12.6.7 implies $\mathbf{y} \in \mathcal{R}(P_\lambda)$. Similarly, $\mathbf{y} \in \mathcal{R}(P_\lambda)$ implies $(\lambda I - A)^{k_\lambda - 2} \mathbf{x} \in \mathcal{R}(P_\lambda)$. Repeating gives $(\lambda I - A)^\ell \mathbf{x} \in \mathcal{R}(P_\lambda)$ for each ℓ . Thus, $\mathcal{E}_\lambda \subset \mathcal{R}(P_\lambda)$.

Finally, we note that $\mathbb{F}^n = \bigoplus \mathcal{R}(P_\lambda)$ and $\mathbb{F}^n = \bigoplus \mathcal{E}_\lambda$. Since $\mathcal{E}_\lambda \subset \mathcal{R}(P_\lambda)$ for each $\lambda \in \sigma(A)$, it follows that $\mathcal{R}(P_\lambda) = \mathcal{E}_\lambda$. $\square$

Remark 12.6.10. The previous theorem holds for any collection of projections and nilpotents satisfying the properties listed in Remark 12.6.8. This is important for the proof that the spectral decomposition is unique (Theorem 12.7.5).

Example 12.6.11. The eigenprojection P_1 of Example 12.6.2 has rank 3, and the eigennilpotent D_1 satisfies

$$D_1^2 = 9 \begin{bmatrix} 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad D_1^3 = 0;$$

hence, D_1 has order 3. On the other hand, P_7 has rank 1 and D_7 has order 1, that is, $D_7 = 0$, as we have already seen.

Theorem 12.6.12 (Spectral Decomposition Theorem). For each $\lambda \in \sigma(A)$ let P_λ denote the spectral projection associated to λ , and let D_λ denote the corresponding eigennilpotent of order m_λ . The resolvent takes the form

$$R(z) = \sum_{\lambda \in \sigma(A)} \left[\frac{P_\lambda}{z - \lambda} + \sum_{k=1}^{m_\lambda-1} \frac{D_\lambda^k}{(z - \lambda)^{k+1}} \right], \quad (12.33)$$

and we have the spectral decomposition

$$A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda. \quad (12.34)$$

Proof. Using (12.32), we have

$$R(z) = R(z) \sum_{\lambda \in \sigma(A)} P_\lambda = \sum_{\lambda \in \sigma(A)} R(z) P_\lambda = \sum_{\lambda \in \sigma(A)} \left[\frac{P_\lambda}{z - \lambda} - \sum_{k=1}^{m_\lambda-1} \frac{D_\lambda^k}{(z - \lambda)^{k+1}} \right].$$

Similarly Lemma 12.6.1 yields

$$A = A \sum_{\lambda \in \sigma(A)} P_\lambda = \sum_{\lambda \in \sigma(A)} A P_\lambda = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda. \quad \square$$

Example 12.6.13. For the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

of Examples 12.4.5, 12.6.2, and 12.6.11 we can show that

$$\sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda = (1P_1 + D_1) + 7P_7 = A.$$

Corollary 12.6.14. Let $f(z)$ be holomorphic in an open, simply connected set containing $\sigma(A)$. For each $\lambda \in \sigma(A)$ let

$$f(z) = f(\lambda) + \sum_{n=1}^{\infty} a_{n,\lambda} (z - \lambda)^n$$

be the Taylor series representation of f near λ . We have

$$f(A) = \sum_{\lambda \in \sigma(A)} \left(f(\lambda) P_\lambda + \sum_{k=1}^{m_\lambda-1} a_{k,\lambda} D_\lambda^k \right). \quad (12.35)$$

In the case that A is semisimple, (12.35) reduces to

$$f(A) = \sum_{\lambda \in \sigma(A)} f(\lambda) P_\lambda. \quad (12.36)$$

Proof. For each $\lambda \in \sigma(A)$, let Γ_λ be a small circle around λ that contains no other points of $\sigma(A)$ and which lies inside of the open, simply connected set U containing $\sigma(A)$. For convenience of notation, set $a_{0,\lambda} = f(\lambda)$. By the spectral resolution formula (Theorem 12.4.6) we have

$$\begin{aligned} f(A) &= \frac{1}{2\pi i} \sum_{\lambda \in \sigma(A)} \oint_{\Gamma_\lambda} f(z) R(z) dz \\ &= \frac{1}{2\pi i} \sum_{\lambda \in \sigma(A)} \oint_{\Gamma_\lambda} \sum_{k=0}^{\infty} a_{k,\lambda} (z - \lambda)^k R(z) dz \\ &= \frac{1}{2\pi i} \sum_{\lambda \in \sigma(A)} \sum_{k=0}^{\infty} a_{k,\lambda} \oint_{\Gamma_\lambda} (z - \lambda)^k R(z) dz \\ &= \sum_{\lambda \in \sigma(A)} \left(f(\lambda) P_\lambda + \sum_{k=1}^{m_\lambda-1} a_{k,\lambda} D_\lambda^k \right). \quad \square \end{aligned}$$

In the next section we show (Theorem 12.7.6) that (12.35) is not just a way to compute $f(A)$, but that it is actually the spectral decomposition of $f(A)$.

Example 12.6.15. For each $\lambda \in \sigma(A)$, let m_λ be the order of D_λ .

(i) Given $k \in \mathbb{N}$, we compute $f(A) = A^k$. The binomial formula gives

$$z^k = (\lambda + (z - \lambda))^k = \sum_{j=0}^k \binom{k}{j} \lambda^{k-j} (z - \lambda)^j, \quad (12.37)$$

which converges everywhere, so if $M_\lambda = \min(k, m_\lambda)$, then

$$A^k = \sum_{\lambda \in \sigma(A)} \left(\lambda^k P_\lambda + \sum_{j=1}^{M_\lambda} \binom{k}{j} \lambda^{k-j} D_\lambda^j \right). \quad (12.38)$$

(ii) Given $t \in \mathbb{R}$, we compute $f(A) = e^{At}$. The Taylor expansion of e^{zt} around λ is

$$e^{zt} = e^{\lambda t} \sum_{k=0}^{\infty} \frac{t^k (z - \lambda)^k}{k!}.$$

This gives the formula

$$e^{At} = \sum_{\lambda \in \sigma(A)} e^{\lambda t} \left(P_\lambda + \sum_{k=1}^{m_\lambda-1} \frac{t^k D_\lambda^k}{k!} \right). \quad (12.39)$$

12.7 Spectral Mapping Theorem

Recall the semisimple spectral mapping theorem (Theorem 4.3.12), which says that for any polynomial $p \in \mathbb{F}[x]$ and any semisimple $A \in M_n(\mathbb{F})$ the set of eigenvalues of $p(A)$ is precisely the set $\{p(\lambda) \mid \lambda \in \sigma(A)\}$. In this section we extend this theorem in two ways. First, the spectral mapping theorem applies to all matrix operators, not just semisimple matrices, and second, it holds for all holomorphic functions, not just polynomials.

The spectral mapping theorem describes the spectrum of $f(A)$, but it can be extended to give a description of the full spectral decomposition of $f(A)$. Specifically, we prove Theorem 12.7.6, which says the spectral decomposition of $f(A)$ has the form described in the previous section; see Corollary 12.6.14. To prove this generalization we first prove Theorem 12.7.5, showing that the spectral decomposition of any matrix operator is unique.

We conclude this section with the *power method* for finding eigenvectors of matrices. This important application of the spectral mapping theorem plays a central role in many applications, including Google's PageRank algorithm.

12.7.1 The Spectral Mapping Theorem

Theorem 12.7.1 (Spectral Mapping Theorem). *For any $A \in M_n(\mathbb{F})$, if $f(z)$ is holomorphic in an open, simply connected set containing $\sigma(A)$, then $\sigma(f(A)) = f(\sigma(A))$. Moreover, if $\mathbf{x}$ is an eigenvector of A corresponding to the eigenvalue λ , then $\mathbf{x}$ is also an eigenvector of $f(A)$ corresponding to $f(\lambda)$.*

Proof. If $\mu \notin f(\sigma(A))$, then $f(z) - \mu$ is both holomorphic in a neighborhood of $\sigma(A)$ and nonzero on $\sigma(A)$. Hence, $f(A) - \mu I$ is nonsingular, or, equivalently, $\mu \notin \sigma(f(A))$.

Conversely, if $\mu \in f(\sigma(A))$, that is $\mu = f(\lambda)$ for some $\lambda \in \sigma(A)$, then set

$$g(z) = \begin{cases} \frac{f(z) - f(\lambda)}{z - \lambda}, & z \neq \lambda, \\ f'(\lambda), & z = \lambda. \end{cases}$$

Note that $g(z)$ is holomorphic in a punctured disk around λ and it is continuous at λ , so by Exercise 11.19 it must be holomorphic in a neighborhood of λ . Also, $g(z)(z - \lambda) = f(z) - \mu$, and hence $g(A)(A - \lambda I) = f(A) - \mu I$.

If $\mathbf{x}$ is an eigenvector of A associated to λ , then

$$(f(A) - \mu I)\mathbf{x} = g(A)(A - \lambda I)\mathbf{x} = g(A)\mathbf{0} = \mathbf{0}.$$

Thus $f(A) - \mu I$ is singular, and $\mu \in \sigma(f(A))$ with $\mathbf{x}$ as an eigenvector. $\square$

Example 12.7.2. Consider the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

from Examples 12.4.5 and 12.6.13, with $\sigma(A) = \{1, 7\}$. The spectral mapping theorem allows us to easily determine the eigenvalues of the following matrices:

- (i) Let $B = A^k$ for $k \in \mathbb{Z}$. Since $f(z) = z^k$ is holomorphic on an open, simply connected set containing $\sigma(A)$, it follows by the spectral mapping theorem that $\sigma(B) = \sigma(A^k) = \{1, 7^k\}$.
- (ii) Let $B = e^{At}$ for $t \in \mathbb{R}$. Since $f(z) = e^{zt}$ is holomorphic on an open, simply connected set containing $\sigma(A)$, it follows by the spectral mapping theorem that

$$\sigma(B) = \sigma(e^{tA}) = e^{t\sigma(A)} = \{e^t, e^{7t}\}.$$

Example 12.7.3. Define $B = \sin(A) + \tan(A^2) + \log(A)$, where A is given in the previous example. Since $f(z) = \sin(z) + \tan(z^2) + \log(z)$ is holomorphic on an open, simply connected set containing $\sigma(A)$, it follows by the spectral mapping theorem that

$$\begin{aligned} \sigma(B) &= \sigma(\sin(A) + \tan(A^2) + \log(A)) \\ &= \sin(\sigma(A)) + \tan(\sigma(A)^2) + \log(\sigma(A)) \\ &= \{\sin(1) + \tan(1), \sin(7) + \tan(49) + \log(7)\}. \end{aligned}$$

Remark 12.7.4. For a less-complicated matrix, like the one found in Example 12.3.4, the spectral mapping theorem may not seem like a big deal, but for more general matrices the spectral mapping theorem can be extremely helpful. The spectral mapping theorem also plays an important role in the proof of the power method for finding eigenvectors and in the proof of the Perron–Frobenius theorem, presented in the next section.

12.7.2 Uniqueness of the Spectral Decomposition

The next theorem shows that there is really only one collection of projections and nilpotents satisfying the main properties of the spectral decomposition. As a simple consequence of this theorem, we show in the next section that for any holomorphic function f and any matrix operator A , the formula for $f(A)$ given in Corollary 12.6.14 also gives the spectral decomposition of $f(A)$.

Theorem 12.7.5. *Given $A \in M_n(\mathbb{F})$, assume that for every $\lambda \in \sigma(A)$ there is a projection $Q_\lambda \in M_n(\mathbb{F})$ and a nilpotent $C_\lambda \in M_n(\mathbb{F})$ satisfying*

- (i) $Q_\lambda^2 = Q_\lambda$,
- (ii) $Q_\lambda Q_\mu = 0$ for all $\mu \in \sigma(A)$ with $\mu \neq \lambda$,
- (iii) $Q_\lambda C_\lambda = C_\lambda Q_\lambda = C_\lambda$,
- (iv) $Q_\mu C_\lambda = C_\lambda Q_\mu = 0$ for all $\mu \in \sigma(A)$ with $\mu \neq \lambda$,
- (v) $\sum_{\lambda \in \sigma(A)} Q_\lambda = I$.

Assume further that

$$A = \sum_{\lambda \in \sigma(A)} \lambda Q_\lambda + C_\lambda. \quad (12.40)$$

In this case, for each $\lambda \in \sigma(A)$, we have

- (i) Q_λ is the eigenprojection P_λ and
- (ii) C_λ is the eigennilpotent D_λ .

Proof. For every $\mu \in \sigma(A)$, the relation (12.40) implies

$$AQ_\mu = \sum_{\lambda \in \sigma(A)} (\lambda Q_\lambda + C_\lambda) Q_\mu = \mu Q_\mu + C_\mu.$$

This gives $C_\mu = (A - \mu I)Q_\mu$. Hence, it suffices to show that $P_\mu = Q_\mu$ for every $\mu \in \sigma(A)$.

Left multiplying (12.40) by Q_λ gives

$$Q_\lambda A = \lambda Q_\lambda + C_\lambda;$$

therefore Remark 12.6.10 shows that Theorem 12.6.9 applies, and $\mathcal{R}(Q_\mu) = \mathcal{E}_\mu$ for every $\mu \in \sigma(A)$. Thus, for any $\mathbf{v} \in V$ we have $P_\mu \mathbf{v} = Q_\mu P_\mu \mathbf{v}$. Hence, for any $\lambda \in \sigma(A)$ we have

$$Q_\lambda \mathbf{v} = \sum_{\mu \in \sigma(A)} Q_\lambda P_\mu \mathbf{v} = Q_\lambda P_\lambda \mathbf{v} = P_\lambda \mathbf{v}.$$

Therefore, $Q_\lambda = P_\lambda$ for all $\lambda \in \sigma(A)$. $\square$

The following theorem is an easy corollary of the uniqueness of the spectral decomposition. It explicitly gives the spectral decomposition of $f(A)$ in terms of the spectral decomposition of A .

Theorem 12.7.6 (Mapping the Spectral Decomposition). *Let $A \in M_n(\mathbb{F})$ and $f(z)$ be holomorphic on an open, simply connected set U containing $\sigma(A)$. For each $\lambda \in \sigma(A)$ let $f(z) = \sum_{n=0}^{\infty} a_{n,\lambda}(z - \lambda)^n$ be the Taylor series representation of f near λ (with $a_{0,\lambda} = f(\lambda)$). The expression*

$$f(A) = \sum_{\lambda \in \sigma(A)} \left(f(\lambda) P_\lambda + \sum_{k=1}^{m_\lambda-1} a_{k,\lambda} D_\lambda^k \right)$$

given in Corollary 12.6.14 is the spectral decomposition of $f(A)$. That is, for each $\nu \in \sigma(f(A))$, the eigenprojection P_ν for $f(A)$ is given by

$$P_\nu = \sum_{\substack{\mu \in \sigma(A) \\ f(\mu) = \nu}} P_\mu \quad (12.41)$$

and the corresponding eigennilpotent D_ν is given by

$$\sum_{\substack{\mu \in \sigma(A) \\ f(\mu) = \nu}} \sum_{k=1}^{m_\mu} a_{k,\mu} D_\mu^k. \quad (12.42)$$

Proof. By the spectral mapping theorem the spectrum of $f(A)$ is $f(\sigma(A))$. Corollary 12.6.14 says that the projections P_ν defined by (12.41) and the nilpotents D_ν defined by (12.42) satisfy (12.40). The other conditions of Theorem 12.7.5 clearly hold; hence these must be the unique eigenprojections and eigennilpotents, respectively, of $f(A)$. $\square$

Example 12.7.7. Theorem 12.7.6 gives the spectral decomposition of A^{-1} as

$$A^{-1} = \sum_{\lambda \in \sigma(A)} \left(\frac{1}{\lambda} P_\lambda + \sum_{\ell=1}^{m_\lambda-1} \frac{(-1)^\ell D_\lambda^\ell}{\lambda^{\ell+1}} \right). \quad (12.43)$$

Thus, if A is the matrix from Examples 12.4.5 and 12.6.13, then we have that

$$A^{-1} = P_1 + (-1)D_1 + (1)D_1^2 + \frac{1}{7}P_7, \quad (12.44)$$

so the 1-eigennilpotent of A^{-1} is $-D_1 + D_1^2$ and the $\frac{1}{7}$ -eigennilpotent is 0. It is easy to check that the decomposition (12.44) adds up to give

$$A^{-1} = \frac{1}{7} \begin{bmatrix} 7 & -21 & 63 & -27 \\ 0 & 7 & -21 & 9 \\ 0 & 0 & 7 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

12.7.3 The Power Method

The power method is another important application of the spectral decomposition formula. If an eigenvalue has its generalized eigenspace equal to its eigenspace—that is, its algebraic and geometric multiplicity are the same—we say it is a *semisimple eigenvalue*. If an operator has a semisimple eigenvalue whose modulus is larger than all the others, the power method gives a way to compute an eigenvector of that dominant eigenvalue.

More precisely, if $A \in M_n(\mathbb{F})$ and $\lambda \in \sigma(A)$ is semisimple, then $\mathcal{E}_\lambda$ is the eigenspace of λ and Theorem 12.6.9 guarantees that $\mathcal{R}(P_\lambda) = \mathcal{E}_\lambda$. Therefore, if $\mathbf{x} \in \mathbb{F}^n$ satisfies $P_\lambda \mathbf{x} \neq \mathbf{0}$, then $P_\lambda \mathbf{x}$ is an eigenvector of A associated to λ . If λ is a dominant semisimple eigenvalue, the power method gives a way to compute the eigenvector $P_\lambda \mathbf{x}$.

Here we prove the power method for a dominant semisimple eigenvalue of modulus 1. Exercise 12.34 extends this to dominant semisimple eigenvalues with arbitrary modulus.

Theorem 12.7.8. *For $A \in M_n(\mathbb{F})$, assume that $1 \in \sigma(A)$ is a semisimple eigenvalue and that there exists $0 \leq \eta < 1$ such that $|\lambda| < \eta$ for all other eigenvalues in $\sigma(A)$. If P is the eigenprojection corresponding to $\lambda = 1$ and $\mathbf{v} \in \mathbb{F}^n$ satisfies $P\mathbf{v} \neq \mathbf{0}$, then as $k \rightarrow \infty$ we have*

$$\|A^k \mathbf{v} - P\mathbf{v}\| \rightarrow 0.$$

Proof. By the spectral resolution theorem (Theorem 12.4.6), we have

$$A^k = \frac{1}{2\pi i} \oint_{\Gamma} z^k R(z) dz,$$

where Γ is a positively oriented circular contour centered at zero having radius greater than 1. Using the residue theorem (Theorem 11.7.13), we have that

$$A^k = \frac{1}{2\pi i} \oint_{\Gamma_\eta} z^k R(z) dz + \frac{1}{2\pi i} \oint_{\Gamma_1} z^k R(z) dz, \quad (12.45)$$

where Γ_η is a positively oriented circular contour centered at zero and having radius η , and Γ_1 is a small positively oriented circular contour centered at 1 and having radius $1 - \eta$.

From (12.37), we have that

$$\frac{1}{2\pi i} \oint_{\Gamma_1} z^k R(z) dz = \frac{1}{2\pi i} \sum_{\ell=0}^k \binom{k}{\ell} \oint_{\Gamma_1} (z-1)^\ell R(z) dz = \sum_{\ell=0}^k \binom{k}{\ell} A_{-\ell-1} = P.$$

The last equality follows because the eigenvalue $\lambda = 1$ is semisimple, and thus $A_{-1} = P$ and $A_{-\ell-1} = 0$ for each $\ell \geq 1$. Combining this with (12.45) we have

$$A^k - P = \frac{1}{2\pi i} \oint_{\Gamma_\eta} z^k R(z) dz.$$

For any given operator norm $\|\cdot\|$, following the proof of Corollary 12.4.7 gives

$$\|A^k - P\| = \left\| \frac{1}{2\pi i} \oint_{\Gamma_\eta} z^k R(z) dz \right\| \leq \frac{1}{2\pi} 2\pi C \eta^k = C \eta^k,$$

where $C = \sup_{z \in \Gamma_\eta} \|R(z)\|$. Thus, as $k \rightarrow \infty$ we have

$$\|A^k - P\| \leq C \eta^k \rightarrow 0. \quad \square$$

Example 12.7.9. Consider the matrix

$$A = \begin{bmatrix} 1/5 & 2/5 & 2/5 \\ 3/5 & 1/5 & 2/5 \\ 1/5 & 2/5 & 1/5 \end{bmatrix}.$$

It is easy to show that $\sigma(A) = \{1, -1/5\}$, with the second eigenvalue having algebraic multiplicity two. Therefore, the eigenvalue 1 is simple. For any $\mathbf{x}_0$ with $P\mathbf{x}_0 \neq \mathbf{0}$, semisimplicity of the eigenvalue $\lambda = 1$ means that $P\mathbf{x}_0$ is an eigenvector of A with eigenvalue 1.

The iterative map $\mathbf{x}_k = A\mathbf{x}_{k-1}$ gives

$$\|\mathbf{x}_k - P\mathbf{x}_0\| = \|A^k\mathbf{x}_0 - P\mathbf{x}_0\| \leq \|A^k - P\|\|\mathbf{x}_0\| \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

Hence, the eigenvector $P\mathbf{x}_0$ is given by $\lim_{k \rightarrow \infty} \mathbf{x}_k$.

After ten steps of applying this method to $\mathbf{x}_0 = [1 \ 1 \ 1]^T$ in double-precision floating-point arithmetic, the sequence stabilizes to

$$\mathbf{x}_{10} = [1.0 \ 1.16666667 \ 0.83333333]^T,$$

matching the correct answer $P\mathbf{x}_0 = [1 \ 7/6 \ 5/6]^T$ (essentially) perfectly.

12.8 The Perron–Frobenius Theorem

In this section we consider matrix operators whose entries are all nonnegative or all positive. Nonnegative and positive matrices arise in applications such as Markov chains in probability theory, compartmental models in differential equations, and Google's PageRank algorithm for information retrieval. In all these applications, the spectral properties of the matrices are important. Some of the most important spectral information for these applications comes from the theorems of Perron and Frobenius, which we describe here.

12.8.1 Perron's Theorem

Definition 12.8.1. A matrix $A \in M_n(\mathbb{R})$ is called nonnegative, denoted $A \succeq 0$, if every entry is nonnegative. It is called positive, denoted $A \succ 0$, if every entry is positive.

Remark 12.8.2. Sometimes it can also be useful to use the notation $B \succeq A$ when $B - A \succeq 0$, and $B \succ A$ when $B - A \succ 0$.

Remark 12.8.3. If $A \succ 0$, then $A^k \succ 0$ for all $k \in \mathbb{Z}^+$, simply because every entry is a sum of products of strictly positive numbers. This implies that A is not nilpotent, and therefore, by Lemma 12.6.3, the spectral radius $r(A)$ must be positive.

Theorem 12.8.4. *A nonnegative matrix $A \succeq 0$ has an eigenvalue λ equal to its spectral radius $r(A)$, and λ has a nonnegative eigenvector.*

Proof. By Theorems 12.3.8 and 12.3.14 the Laurent series

$$R(z) = \frac{1}{z}I + \frac{1}{z^2}A + \frac{1}{z^3}A^2 + \cdots \quad (12.46)$$

converges for all z with $|z| > r(A)$.

By way of contradiction, assume that $\lambda = r(A)$ is not an eigenvalue of A . This implies the value of $\lim_{z \rightarrow \lambda} R(z)$ is finite, so the series

$$R(\lambda) = \frac{1}{\lambda}I + \frac{1}{\lambda^2}A + \frac{1}{\lambda^3}A^2 + \cdots$$

converges as well. Since $A^n \succeq 0$ for every $n > 0$, the convergence is componentwise absolute. Writing $z = 1/w$, the power series $w + w^2A + w^3A^2 + \cdots$ converges uniformly on the compact disk $\overline{B(0, 1/\lambda)}$, and hence $R(z) = z^{-1} + z^{-2}A + z^{-3}A^2 + \cdots$ converges uniformly on the closed set $\{|z| \geq \lambda\}$. Thus the resolvent set includes the entire circle $|z| = \lambda$, and no eigenvalue lies on the circle. Hence $r(A) < \lambda$, a contradiction.

To see that a nonnegative eigenvector of λ exists, note that by (12.33) we have

$$R(z)D_\lambda^{m_\lambda-1} = \sum_{\mu \in \sigma(A)} \left[\frac{P_\mu}{z - \mu} + \sum_{k=1}^{m_\mu-1} \frac{D_\mu^k}{(z - \mu)^{k+1}} \right] D_\lambda^{m_\lambda-1} = \frac{D_\lambda^{m_\lambda-1}}{z - \lambda}, \quad (12.47)$$

where m_λ is the order of the eigennilpotent D_λ (define $D_\lambda^{m_\lambda-1} = P_\lambda$ if $m_\lambda = 1$). But $R(z) = (zI - A)^{-1}$, so (12.47) implies that

$$(zI - A)D_\lambda^{m_\lambda-1} = (z - \lambda)D_\lambda^{m_\lambda-1},$$

which gives

$$AD_\lambda^{m_\lambda-1} = \lambda D_\lambda^{m_\lambda-1}. \quad (12.48)$$

Hence, the nonzero columns of $D_\lambda^{m_\lambda-1}$ are eigenvectors of λ . By (12.31) we also have

$$D_\lambda^{m_\lambda-1} = \lim_{z \rightarrow \lambda} (z - \lambda)^{m_\lambda} R(z).$$

But for $|z| > \lambda = r(A)$, the resolvent $R(z)$ can be written (see Theorem 12.3.8) as $R(z) = \sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}}$. Since $A \succeq 0$ we also have $\lim_{z \rightarrow \lambda+} (z - \lambda)^{m_\lambda} R(z) \succeq 0$. $\square$

The next theorem shows that for positive matrices the real eigenvalue equal to its spectral radius is simple and the eigenvector is componentwise positive. We will need the following simple lemma in the proof.

Lemma 12.8.5. *If $B \in M_n(\mathbb{R})$ is nonnegative and has a nonzero entry on the diagonal, then B is not nilpotent.*

Proof. Let $B, C \in M_n(\mathbb{R})$ with $B, C \succeq 0$. Assume there exists some k such that $b_{kk} > 0$ and $c_{kk} > 0$. The (k, k) entry of BC is

$$\sum_{i=1}^n b_{ki}c_{ik} = b_{kk}c_{kk} + \sum_{i \neq k} b_{ki}c_{ik}, \quad (12.49)$$

but $b_{kk}c_{kk}$ is strictly positive, and the remaining terms are all nonnegative, so the (k, k) entry of BC is strictly positive.

If $b_{kk} > 0$, then by induction, using (12.49) with $C = B^{m-1}$, the (k, k) entry of $B^m = B B^{m-1}$ is also strictly positive for all $m \in \mathbb{Z}^+$; hence $B^m \neq 0$. $\square$

Theorem 12.8.6 (Perron). *A positive matrix $A \succ 0$ has a simple (multiplicity one) eigenvalue λ equal to its spectral radius $r(A)$. In addition, λ has a positive eigenvector. Moreover all other eigenvalues of A are smaller in modulus than $r(A)$.*

Proof. Assume $A \succ 0$. Let $\mathbf{v}_1$ be a nonnegative eigenvector belonging to the eigenvalue $\lambda = r(A)$. Note that $A\mathbf{v}_1 \succ \mathbf{0}$, and since $A\mathbf{v}_1 = \lambda\mathbf{v}_1$ and $\lambda > 0$, it follows that $\mathbf{v}_1 \succ \mathbf{0}$ as well. From (12.48), we know each nonzero column of $D_\lambda^{m_\lambda-1}$ is a nonnegative eigenvector, and hence is positive. By Lemma 12.8.5 this contradicts the nilpotency of D_λ . Hence, $D_\lambda = 0$, and the generalized eigenspace $\mathcal{E}_\lambda$ is spanned by eigenvectors of λ .

Since A and λ are both real, the kernel of $(A - \lambda I)$ has a basis of real vectors. Suppose that $\mathbf{v}_2$ is an element of this basis that is linearly independent from $\mathbf{v}_1$, and assume that $\mathbf{v}_2$ has at least one negative entry (if not, multiply by -1). For every t , the vector $t\mathbf{v}_1 + \mathbf{v}_2$ is also an eigenvector with eigenvalue λ . Since $\mathbf{v}_1 \succ \mathbf{0}$, there must be some real t such that $t\mathbf{v}_1 + \mathbf{v}_2 \succeq \mathbf{0}$ and such that at least one coordinate of $t\mathbf{v}_1 + \mathbf{v}_2$ has a zero coordinate. This is a contradiction, since nonnegative eigenvectors of positive matrices are positive. Hence, the dimension of the λ eigenspace is 1, and λ is a simple eigenvalue.

Finally, we show that A has no other eigenvalues on the circle $|z| = r(A)$. Choose $\varepsilon > 0$ small enough that $A - \varepsilon I \succeq 0$. By definition of eigenvalue (or by the spectral mapping theorem), $\lambda - \varepsilon$ is the largest positive eigenvalue of $A - \varepsilon I$. It follows from Theorem 12.8.6 that $r(A - \varepsilon I) = \lambda - \varepsilon$. Again by the definition of eigenvalue we have that $\sigma(A) = \sigma(A - \varepsilon I) + \varepsilon \subset \overline{B(\varepsilon, \lambda - \varepsilon)}$. Thus $\lambda \in \overline{B(\varepsilon, \lambda - \varepsilon)}$ is the only eigenvalue on the circle $|z| = \lambda$; see Figure 12.4. $\square$

Remark 12.8.7. For any nonnegative matrix A , if a is the smallest diagonal entry of A , then the end of the previous proof (taking $\varepsilon = a$) shows that $\sigma(A) \subset \overline{B(a, r(A) - a)}$. Thus, in the case of a nonnegative matrix with positive diagonal, we still have the conclusion that the eigenvalue $\lambda = r(A)$ (often called the *Perron root* or *Perron–Frobenius eigenvalue*) is the only eigenvalue on the circle $|z| = r(A)$.

Example 12.8.8. The positive matrix

$$A = \begin{bmatrix} 1/5 & 2/5 & 2/5 \\ 3/5 & 1/5 & 2/5 \\ 1/5 & 2/5 & 1/5 \end{bmatrix}$$

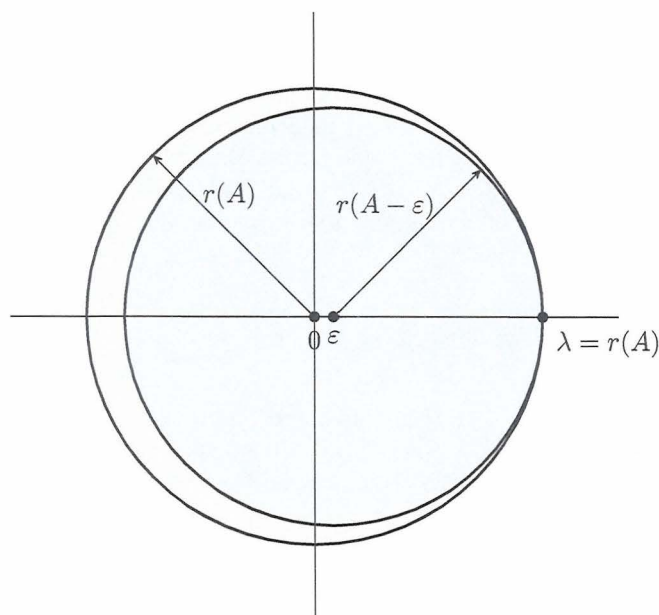


Figure 12.4. Illustration to assist in the proof of Theorem 12.8.6. Every eigenvalue of a positive matrix A lies in $\overline{B(\varepsilon, \lambda - \varepsilon)}$, where ε is chosen small enough that $A - \varepsilon I \succeq 0$. The Perron root $\lambda = r(A)$ is the only point in $\overline{B(\varepsilon, \lambda - \varepsilon)}$ on the circle $|z| = \lambda$.

of Example 12.7.9 has spectrum $\sigma(A) = \{1, -1/5\}$, and the eigenvalue $-1/5$ has multiplicity two, so the eigenvalue 1 (the Perron root) is simple and is equal to the spectral radius $r(A) = 1$. Example 12.7.9 also shows that the corresponding eigenvector is $[1 \ 7/6 \ 5/6]^T$, which is strictly positive, as guaranteed by Perron's theorem.

12.8.2 Perron–Frobenius Theorem

Frobenius extended the uniqueness result of Perron's theorem to a large class of nonnegative matrices called *irreducible*.

Definition 12.8.9. A nonnegative matrix $A \succeq 0$ is called *primitive* if there is a k such that $A^k \succ 0$. We say that A is *irreducible* if for any i, j there is a $k \in \mathbb{Z}^+$ such that the (i, j) entry of A^k is positive.

Proposition 12.8.10. If $A \succeq 0$ is irreducible, then $I + A$ is primitive.

Proof. The proof is Exercise 12.38. $\square$

Theorem 12.8.11 (Perron–Frobenius). A nonnegative irreducible matrix $A \succeq 0$ has a simple eigenvalue λ equal to its spectral radius $r(A)$, and λ has a positive eigenvector.

Proof. We know from Theorem 12.8.4 that A has an eigenvalue λ equal to its spectral radius $r(A)$ and that λ has a nonnegative eigenvector. What remains to be shown is that λ is simple and the nonnegative eigenvector is positive. By Proposition 12.8.10 the matrix $(I + A)$ is primitive. We define $B = (I + A)^k$, where $k \in \mathbb{Z}^+$ is chosen to be large enough so that $B \succ 0$. It follows from the spectral mapping theorem that $\lambda \in \sigma(A)$ if and only if $(1 + \lambda)^k \in \sigma(B)$, and thus the algebraic multiplicity of λ is equal to the algebraic multiplicity of $(1 + \lambda)^k$. Observe that

$$r(B) = \max_{\lambda \in \sigma(A)} |(1 + \lambda)^k| = \max_{\lambda \in \sigma(A)} |(1 + \lambda)|^k = \left\{ \max_{\lambda \in \sigma(A)} |(1 + \lambda)| \right\}^k = (1 + r(A))^k$$

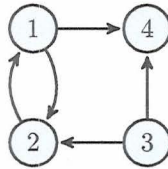
because when the disk $|z| \leq r(A)$ is translated one unit to the right, the point of maximum modulus is $z = 1 + r(A)$. Since B is a positive matrix, the algebraic multiplicity of the eigenvalue at $r(B)$ is one, and so the algebraic multiplicity of λ must also be one.

Finally, let $\mathbf{v}$ be the nonnegative eigenvector of A corresponding to $\lambda = r(A)$. Since $\mathbf{v} = [v_1, \dots, v_n]^T$ is not identically zero, there exists an i with $v_i \neq 0$. For each $j \in \{1, \dots, n\}$ there exists a k such that the (i, j) entry of A^k is positive (since A is irreducible). This implies that the i th entry of $A^k \mathbf{v}$ is positive. But we have $A^k \mathbf{v} = \lambda^k \mathbf{v}$, and thus $v_i > 0$. Since this holds for every j , we have $\mathbf{v} \succ \mathbf{0}$. $\square$

12.8.3 Google's PageRank Algorithm

Google searches are useful because they not only return a list of web pages containing your results, but they also rank the results. The *PageRank* algorithm is one of the key tools used to determine the rank of each website.⁴⁷

The basic idea of the algorithm is to simulate web traffic mathematically and then rank pages by the percentage of traffic they get. Consider the following toy model of the Internet, consisting of only four pages. The arrows in the diagram correspond to the links from one page to another.



Assume that at each time step a user moves to a new page by clicking on a link (selected with equal likelihood). Let $A = [a_{ij}] \in M_4(\mathbb{R})$, where a_{ij} is the probability that a user at page j will click on the link for page i . At the first page a user will click on pages two or four with equal probability; hence, $a_{21} = a_{41} = 0.5$. A user at page two will click on page one with probability one; hence, $a_{12} = 1$. The third page has two links, going to pages two and four. Thus, $a_{23} = a_{43} = 0.5$.

⁴⁷This algorithm is named after Larry Page, one of Google's founders.

The fourth page presents a problem, because it has no outbound links. Instead of setting all the corresponding entries to 0, we assume the user will randomly “teleport” to another page (all with equal probability), so $a_{i4} = 0.25$ for each i . Putting this all together, we have

$$A = \begin{bmatrix} 0 & 1 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0.25 \end{bmatrix}.$$

If $\mathbf{e}_k \in \mathbb{R}^4$ is the k th standard basis vector, then $A\mathbf{e}_k$ is the k th column of A —the vector describing the probability that a user starting at page k will move to each of the other pages. If the k th entry of $\mathbf{x} \in \mathbb{R}^4$ is the percentage of all users currently on page k , then the product $A\mathbf{x}$ describes the expected percentage of users that will be on each of the pages after the next step. Repeating the process, $A^2\mathbf{x}$ describes the percentage of traffic that will be at each page after two steps, and $A^k\mathbf{x}$ the percentage of traffic at each page after k steps.

Notice that A is nonnegative and every column sums to 1, so A has a left eigenvector $\ell = \mathbf{1}^\top$ of all ones, with corresponding eigenvalue 1. But we also have $\|A\|_1 = 1$, so Lemma 12.3.13 implies that $r(A) \leq 1$; hence $r(A) = 1$.

A right eigenvector $\mathbf{r}$ corresponding to the eigenvalue 1 satisfies $A\mathbf{r} = \mathbf{r}$. If $\mathbf{r}$ is scaled so that its entries are nonnegative and sum to 1, then it represents the distribution of traffic in a steady state, where the overall distribution of traffic at the next time step $A\mathbf{r}$ is the same as the distribution $\mathbf{r}$ at the current time. If the eigenvalue 1 has a one-dimensional eigenspace, then there is a unique nonnegative choice of $\mathbf{r}$ whose entries sum to 1, and the percentage r_k of traffic at the k th page is a reasonable indicator of how important that page is.

The same logic applies to an arbitrary number n of pages with any configuration of links. Again construct a matrix $A \in M_n(\mathbb{F})$ corresponding to the link probabilities, following the approach described above. Since A is nonnegative and its columns sum to 1, the previous argument shows that $r(A) = 1$, and if the eigenvalue 1 is simple, then there is a unique nonnegative right eigenvector $\mathbf{r}$ whose entries sum to 1. This eigenvector gives the desired ranking of pages.

The one remaining problem is that the eigenvalue 1 is not necessarily simple. To address this, the PageRank algorithm assumes that at each time step a percentage $\varepsilon < 1$ of users follow links in A and the remaining percentage $1 - \varepsilon > 0$ teleport to a random web page. If all web pages are equally likely for those who teleport, the new matrix of page-hit probabilities is given by

$$B = \varepsilon A + \frac{1 - \varepsilon}{n} E, \quad (12.50)$$

where $E \in M_n(\mathbb{R})$ is the matrix of all ones, that is, $E = \mathbf{1}\mathbf{1}^\top$. All of the columns of B sum to one, and applying the same argument used above gives $r(B) = 1$.

Depending on the choice of ε , the matrix B might be a more realistic model for Internet traffic than A is, since some users really do just move to a new page without following a link. But another important advantage of B over A is that $B \succ 0$, so Perron’s theorem (Theorem 12.8.6) applies, guaranteeing the eigenvalue 1 is simple. Thus, there is a unique positive right eigenvector $\mathbf{r}$ whose entries sum to 1, and we can use r to rank the importance of web pages. Moreover,

since the Perron root 1 is simple, the power method (see Theorem 12.7.8) guarantees that for every nontrivial initial choice $\mathbf{x}_0 \succeq \mathbf{0}$, the sequence $\mathbf{x}_0, A\mathbf{x}_0, A^2\mathbf{x}_0, \dots$ converges to $\mathbf{r}$.

Vista 12.8.12. Probability matrices like those considered above are examples of *Markov chains*, which have widespread applications. Both Perron's theorem and the power method are key tools in the study of Markov chains.

12.9 The Drazin Inverse

The Moore–Penrose inverse $A^\dagger$ (defined in Theorem 4.6.1) of a square matrix A satisfies $AA^\dagger = \text{proj}_{\mathcal{R}(A)}$ and $A^\dagger A = \text{proj}_{\mathcal{R}(A^H)}$. In other words, the products $AA^\dagger$ and $A^\dagger A$ are geometrically the nearest possible approximations to I . But in settings where we are interested in multiplicative and spectral properties, rather than geometric properties, the Moore–Penrose inverse is not usually the best choice.

In this section we describe a different generalized inverse, called the *Drazin inverse*, that behaves much better with respect to spectral and multiplicative properties. In some sense, the only obstruction to inverting a matrix is the existence of zero eigenvalues. Thus, instead of using a projection to the range of A or A^H , as in the Moore–Penrose inverse, the Drazin inverse uses a projection to the generalized eigenspaces associated to nonzero eigenvalues.

The Drazin inverse has many applications in dynamical systems, Markov chains, and control theory, among others. It also is an important theoretical concept in the study of Krylov subspaces and the corresponding results in numerical linear algebra. In the next chapter we show how the Drazin inverse connects to the Arnoldi and GMRES algorithms, which are used to solve large sparse linear algebra problems.

12.9.1 Definition and Spectral Decomposition

Definition 12.9.1. For $A \in M_n(\mathbb{F})$ define $P_* = I - P_0$, where P_0 is the 0-eigenprojection in the spectral decomposition

$$A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda. \quad (12.51)$$

The projections P_0 and P_* commute with A , and so $\mathcal{R}(P_0)$ and $\mathcal{R}(P_*)$ are both A -invariant. We can write $\mathbb{F}^n = \mathcal{R}(P_0) \oplus \mathcal{R}(P_*)$, and since $I = P_0 + P_*$ we have

$$A = AP_0 + AP_* = D_0 + AP_*. \quad (12.52)$$

The decomposition (12.52) is sometimes called the *Wedderburn decomposition* of A .

Definition 12.9.2. Let $A \in M_n(\mathbb{F})$ have spectral decomposition (12.51), and let P_* be as in Definition 12.9.1. Let C denote the restriction of A to $\mathcal{R}(P_*)$. The operator C has no zero eigenvalues, so it is invertible. Define the Drazin inverse A^D of A to be the operator

$$A^D = C^{-1} \circ P_*.$$

Alternatively, if $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_* \in \mathcal{R}(P_0) \oplus \mathcal{R}(P_*)$, then $A^D(\mathbf{x}) = C^{-1}\mathbf{x}_* \in \mathcal{R}(P_*)$.

Remark 12.9.3. When A is invertible, then $P_0 = 0$ and $P_* = I$, so we have $A^D = A^{-1}$.

Remark 12.9.4. Since $\mathbb{F}^n = \mathcal{R}(P_*) \oplus \mathcal{N}(P_*)$, we may choose a basis such that the operator A can be written in block form as

$$A = S^{-1} \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} S, \quad (12.53)$$

where S is the change of basis matrix that block diagonalizes A into complementary subspaces, N is a nilpotent block of order k , and the block M is the matrix representation of C in this basis. Thus, we have

$$AP_* = S^{-1} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} S, \quad D_0 = S^{-1} \begin{bmatrix} 0 & 0 \\ 0 & N \end{bmatrix} S, \quad \text{and} \quad A^D = S^{-1} \begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} S.$$

The spectral decomposition of the Drazin inverse is just like the spectral decomposition for the usual inverse given in (12.43), except that the terms that would have corresponded to the eigenvalue 0 are missing, as shown in the next theorem.

Theorem 12.9.5. For $A \in M_n(\mathbb{F})$ with spectral decomposition (12.51), the Drazin inverse A^D can be written as

$$A^D = \sum_{\substack{\lambda \in \sigma(A) \\ \lambda \neq 0}} \left(\frac{1}{\lambda} P_\lambda + \sum_{\ell=1}^{m_\lambda-1} \frac{(-1)^\ell D_\lambda}{\lambda^{\ell+1}} \right), \quad (12.54)$$

where each m_λ denotes the algebraic multiplicity of the eigenvalue $\lambda \in \sigma(A)$.

Proof. Write the spectral decomposition of C as

$$C = \sum_{\lambda \in \sigma(C)} \lambda P_{C,\lambda} + D_{C,\lambda},$$

and observe that $\sigma(C) = \sigma(A) \setminus \{0\}$. Using (12.43) gives the spectral decomposition of C^{-1} as

$$C^{-1} = \sum_{\lambda \in \sigma(C)} \left(\frac{1}{\lambda} P_{C,\lambda} + \sum_{\ell=1}^{m_\lambda-1} \frac{(-1)^\ell D_{C,\lambda}}{\lambda^{\ell+1}} \right).$$

By the definition of C , we have that $P_{C,\lambda} \circ P_* = P_\lambda$ and $D_{C,\lambda} \circ P_* = D_\lambda$ for every $\lambda \in \sigma(C)$, and this implies that

$$\begin{aligned} A^D &= C^{-1} \circ P_* \\ &= \sum_{\lambda \in \sigma(C)} \left(\frac{1}{\lambda} P_{C,\lambda} + \sum_{\ell=1}^{m_\lambda-1} \frac{(-1)^\ell D_{C,\lambda}}{\lambda^{\ell+1}} \right) \circ P_* \\ &= \sum_{\substack{\lambda \in \sigma(A) \\ \lambda \neq 0}} \left(\frac{1}{\lambda} P_\lambda + \sum_{\ell=1}^{m_\lambda-1} \frac{(-1)^\ell D_\lambda}{\lambda^{\ell+1}} \right). \quad \square \end{aligned}$$

Example 12.9.6. Consider the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The eigenprojection and eigennilpotent for $\lambda = 1$ are

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 27 \\ 0 & 1 & 0 & -9 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad D_1 = \begin{bmatrix} 0 & 3 & 0 & -27 \\ 0 & 0 & 3 & 9 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is easy to check that $A = P_1 + D_1$. By (12.54) the Drazin inverse is $A^D = P_1 - D_1 + D_1^2$, which turns out to be

$$A^D = \begin{bmatrix} 1 & -3 & 9 & 81 \\ 0 & 1 & -3 & -18 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is straightforward to verify that $A^D A = A A^D = P_1$ and $A^D A A^D = A^D$.

Example 12.9.7. For the matrix A from Example 12.9.6 we can use the Wedderburn decomposition to compute A^D and show that we get the same answer as we got from the spectral decomposition in that example. For details, see Exercise 12.43.

12.9.2 Alternative Characterizations

Theorem 12.9.8. *Let $A \in M_n(\mathbb{F})$ be a singular matrix. If Γ is a positively oriented simple closed contour that contains all of the eigenvalues of A except the origin, then*

$$A^D = \frac{1}{2\pi i} \oint_{\Gamma} z^{-1} R(z) dz. \quad (12.55)$$

In other words, (12.55) is an alternative definition of the Drazin inverse.

Proof. Note that the function $f(z) = 1/z$ has Taylor series around $\lambda \neq 0$ equal to

$$\frac{1}{z} = \sum_{\ell=0}^{\infty} \frac{(z - \lambda)^{\ell} (-1)^{\ell}}{\lambda^{\ell+1}}.$$

By the same argument as the proof of Corollary 12.6.14, we have that

$$\frac{1}{2\pi i} \oint_{\Gamma} z^{-1} R(z) dz = \sum_{\substack{\lambda \in \sigma(A) \\ \lambda \neq 0}} \left(\frac{1}{\lambda} P_{\lambda} + \sum_{\ell=1}^{m_{\lambda}-1} \frac{D_{\lambda}^{\ell} (-1)^{\ell}}{\lambda^{\ell+1}} \right) = A^D. \quad \square$$

Our definition of the Drazin inverse is somewhat unusual. It is more traditional, but perhaps less intuitive, to define the Drazin inverse to be the unique operator satisfying three particular properties. These three properties are listed in the following proposition. We prove that these two approaches are equivalent by showing that A^D satisfies the three properties and that these properties uniquely identify A^D .

Proposition 12.9.9. *For $A \in M_n(\mathbb{F})$ with index $\text{ind}(A) = k$, the Drazin inverse A^D of A satisfies*

- (i) $AA^D = A^D A$,
- (ii) $A^{k+1}A^D = A^k$, and
- (iii) $A^D AA^D = A^D$.

Proof. Writing A and A^D as in Remark 12.9.4, we have

$$AA^D = S^{-1} \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} S = S^{-1} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} S = P_*.$$

Similarly, we have

$$A^D A = S^{-1} \begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} S = S^{-1} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} S = P_*.$$

This establishes (i).

Given $k = \text{ind}(A)$, we have

$$A^{k+1}A^D = A^k(AA^D) = A^k P_* = A^k(I - P_0) = A^k - A^k P_0 = A^k,$$

since $A^k P_0 = 0$. This establishes (iii). Finally, since $A^D P_0 = 0$, we have

$$A^D A A^D = A^D P_* = A^D (I - P_0) = A^D + A^D P_0 = A^D.$$

This establishes (ii). $\square$

Proposition 12.9.10. *The three properties of the previous proposition uniquely determine A^D .*

Proof. Suppose there exists $B \in M_n(\mathbb{F})$ such that

- (i) $AB = BA$,
- (ii) $A^{k+1}B = A^k$, and
- (iii) $BAB = B$.

Since $I = P_* + P_0$, it suffices to show that $BP_* = A^D P_*$ and $BP_0 = A^D P_0 = 0$.

From (i) and the fact that P_0 and P_* commute with A , we have

$$AP_*BP_0 = P_*ABP_0 = P_*BAP_0 = P_*BP_0A,$$

and thus

$$AP_*BP_0A^{k-1} = P_*BP_0A^k = 0.$$

Since A is invertible on $\mathcal{R}(P_*)$, this gives

$$P_*BP_0A^{k-1} = 0.$$

Using the argument again, we have

$$AP_*BP_0A^{k-2} = P_*BP_0A^{k-1} = 0,$$

which yields

$$P_*BP_0A^{k-2} = 0.$$

Continuing inductively gives

$$P_*BP_0 = 0 \quad \text{and} \quad BP_0 = P_0BP_0.$$

A similar argument gives

$$P_0BP_* = 0 \quad \text{and} \quad BP_* = P_*BP_*.$$

Combining these results gives

$$P_*B = P_*(BP_0 + BP_*) = P_*BP_* = BP_* \quad \text{and} \quad P_0B = BP_0.$$

By (ii) we have

$$A^{k+1}P_*B = A^{k+1}BP_* = A^kP_* = A^{k+1}A^D P_* = A^{k+1}P_*A^D.$$

This implies that

$$BP_* = P_*B = P_*A^D = A^D P_* = A^D,$$

since A is invertible on $\mathcal{R}(P_*)$.

Finally, (iii) gives

$$BP_0 = P_0B = B^2P_0A.$$

Multiplying on the right by A^{k-1} gives

$$BP_0A^{k-1} = B^2P_0A^k = 0.$$

Continuing inductively, we find

$$BP_0 = 0,$$

as required. $\square$

12.9.3 Application to Differential Equations

Consider the first-order linear differential equation

$$A\mathbf{x}'(t) = B\mathbf{x}(t), \quad A, B \in M_n(\mathbb{C}), \mathbf{x} \in C^1(\mathbb{R}, \mathbb{F}^n). \quad (12.56)$$

If A is invertible, then the general solution is given by $\mathbf{x}(t) = e^{A^{-1}Bt}\mathbf{q}$, where $\mathbf{q} \in \mathbb{F}^n$ is arbitrary. However, if A is singular, then the general solution is more complicated.

Assume that there exists $\lambda \in \mathbb{F}$ such that $(\lambda A + B)^{-1}$ exists. Let

$$\hat{A} = (\lambda A + B)^{-1}A \quad \text{and} \quad \hat{B} = (\lambda A + B)^{-1}B.$$

Since $\lambda\hat{A} + \hat{B} = I$, we have

$$\hat{A}\hat{B} = \hat{A}(I - \lambda\hat{A}) = (I - \lambda\hat{A})\hat{A} = \hat{B}\hat{A}.$$

Multiplying both sides of (12.56) by $(\lambda A + B)^{-1}$, we rewrite the system as

$$\hat{A}\mathbf{x}'(t) = \hat{B}\mathbf{x}(t). \quad (12.57)$$

Let $P_* = \hat{A}^D\hat{A} = \hat{A}\hat{A}^D$ be the projection of Definition 12.9.1 for the matrix $\hat{A}$, and let P_0 and D_0 be the 0-eigenprojection and 0-eigennilpotent, respectively, of $\hat{A}$. We show below that the general solution is $\mathbf{x}(t) = e^{\hat{A}^D\hat{B}t}P_*\mathbf{q}$, where $\mathbf{q} \in \mathbb{F}^n$ is arbitrary.

Taking the derivative of the proposed solution, we have that $\mathbf{x}'(t) = \hat{A}^D\hat{B}\mathbf{x}(t)$. Plugging it into (12.56) gives $P_*\hat{B}\mathbf{x}(t) = \hat{B}\mathbf{x}(t)$. Since $\hat{B}$ commutes with P_* , it suffices to show that $P_0\mathbf{x}(t) \equiv \mathbf{0}$.

Recall that $\hat{B} = I - \lambda\hat{A}$. Multiplying (12.57) by P_0 gives

$$D_0\mathbf{x}'(t) = (P_0 - \lambda D_0)\mathbf{x}(t). \quad (12.58)$$

Multiplying this by D_0^{k-1} gives $D_0^{k-1}\mathbf{x}(t) \equiv \mathbf{0}$, since $D_0^k = 0$. Differentiating gives $D_0^{k-1}\mathbf{x}'(t) \equiv \mathbf{0}$. Thus, multiplying (12.58) by D_0^{k-2} gives $D_0^{k-2}\mathbf{x}(t) \equiv \mathbf{0}$, because the other two terms vanish from the previous case. Repeating reduces the exponent on D_0 until we eventually have the desired result $P_0\mathbf{x}(t) \equiv \mathbf{0}$.

12.10 * Jordan Canonical Form

In Sections 12.5 and 12.6, we showed that a matrix operator $A \in M_n(\mathbb{F})$ can be written as a sum $A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda$, where each P_λ and D_λ are, respectively, the eigenprojections and eigennilpotents corresponding to the eigenvalues $\lambda \in \sigma(A)$. Recall that P_λ and D_λ can be determined by taking contour integrals in the complex plane; see (12.18) and (12.29) for details. This is a *basis-free* approach to spectral theory, meaning that the spectral decomposition can be described abstractly (as contour integrals) without referencing a specific choice of basis. In this framework, we can also choose to use whatever basis is most natural or convenient for the application at hand, and we are free to change to a different basis as needed, and still preserve the decomposition.

In contrast, the development of spectral theory in most texts requires an operator to be represented in a very specific basis, called the *Jordan basis*, which gives an aesthetically pleasing representation of the spectral decomposition. In this section, we describe the Jordan basis and the corresponding matrix representation called the *Jordan canonical form* or *Jordan normal form*, which is a very-nearly-diagonal matrix with zeros everywhere except possibly on the diagonal and the superdiagonal. The diagonal elements are the eigenvalues, and the superdiagonal consists of only zeros and ones.

Unfortunately, the Jordan form is poorly conditioned and thus causes problems with numerical computation; that is, small errors in the floating-point representation of a matrix can compound into large errors in the final results. For this reason, and also because choosing a specific basis is often unnecessarily constraining, the basis-free methods discussed earlier in this chapter are almost always preferable to the Jordan form, but there are occasional times where the basis-specific method can be useful.

12.10.1 Nilpotent Operators

To begin we need a definition and a few results about matrix representations of nilpotent operators.

Definition 12.10.1. For $A \in M_n(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$, the k th Krylov subspace of A generated by $\mathbf{b}$ is

$$\mathcal{K}_k(A, \mathbf{b}) = \text{span}\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{k-1}\mathbf{b}\}.$$

If $\dim(\mathcal{K}_k(A, \mathbf{b})) = k$, we call $\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{k-1}\mathbf{b}\}$ the Krylov basis of $\mathcal{K}_k(A, \mathbf{b})$.

Proposition 12.10.2. Let N be a nilpotent operator on a vector space V , and let $\mathbf{x} \in V$ be such that $N^m(\mathbf{x}) = \mathbf{0}$, but $N^{m-1}(\mathbf{x}) \neq \mathbf{0}$. Expressed in terms of the Krylov basis $\{\mathbf{x}, N(\mathbf{x}), \dots, N^{m-2}(\mathbf{x}), N^{m-1}(\mathbf{x})\}$, the restriction D of N to the Krylov subspace $\mathcal{K}_m(N, \mathbf{x})$ has matrix representation

$$D = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}. \quad (12.59)$$

Proof. See Proposition 12.2.7 and Exercise 12.50. $\square$

Proposition 12.10.3. *Let D be a nilpotent operator on an n -dimensional vector space V . There exist elements $\mathbf{x}_1, \dots, \mathbf{x}_\ell \in V$ and nonnegative integers $d_1, \dots, d_\ell$ such that $n = \sum_{i=1}^\ell d_i$ and V is the direct sum of the Krylov subspaces*

$$V = \bigoplus_{i=1}^\ell \mathcal{K}_{d_i}(D, \mathbf{x}_i).$$

Proof. The proof is by induction on the dimension of V , and it is vacuously true if $\dim(V) = 0$. Assume now that $\dim(V) > 0$. If $\mathcal{R}(D) = \{\mathbf{0}\}$, then taking $\mathbf{x}_1, \dots, \mathbf{x}_\ell$ to be any basis of V and all $d_i = 1$ gives the desired result. We may assume, therefore, that $\mathcal{R}(D)$ contains at least one nonzero element.

Nilpotence of D implies that $\mathcal{R}(D) \neq V$ and hence $\dim(\mathcal{R}(D)) < \dim(V)$. By the induction hypothesis there exist $\mathbf{y}_1, \dots, \mathbf{y}_k \in \mathcal{R}(D)$ and $n_1, \dots, n_k \in \mathbb{N}$ such that

$$\mathcal{R}(D) = \bigoplus_{i=1}^k \mathcal{K}_{n_i}(D, \mathbf{y}_i). \quad (12.60)$$

For each i let $d_i = n_i + 1$ and choose $\mathbf{x}_i \in V$ so that $D\mathbf{x}_i = \mathbf{y}_i$ (this exists because $\mathbf{y}_i \in \mathcal{R}(D)$). The elements $D^{n_1}\mathbf{x}_1, \dots, D^{n_k}\mathbf{x}_k$ all lie in $\mathcal{N}(D)$. Moreover, by (12.60), they are all linearly independent and span $\mathcal{N}(D) \cap \mathcal{R}(D)$, and by the replacement theorem (Theorem 1.4.1) we may choose $\mathbf{x}_{k+1}, \dots, \mathbf{x}_\ell$ in $\mathcal{N}(D)$ so that $D^{n_1}\mathbf{x}_1, \dots, D^{n_k}\mathbf{x}_k, \mathbf{x}_{k+1}, \dots, \mathbf{x}_\ell$ form a basis for $\mathcal{N}(D)$ and $d_{k+1} = \dots = d_\ell = 1$.

If there exists any nontrivial linear combination of the elements

$$\mathbf{x}_1, D\mathbf{x}_1, \dots, D^{d_1-1}\mathbf{x}_1, \dots, \mathbf{x}_\ell, D\mathbf{x}_\ell, \dots, D^{d_\ell-1}\mathbf{x}_\ell,$$

equaling zero, then applying D to the sum and using (12.60) gives a contradiction. Hence these elements must be linearly independent and

$$\sum_{i=1}^\ell \mathcal{K}_{d_i}(D, \mathbf{x}_i) = \bigoplus_{i=1}^\ell \mathcal{K}_{d_i}(D, \mathbf{x}_i) \subset V.$$

Since $\text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k) \cap \mathcal{R}(D) = \text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k) \cap \mathcal{N}(D) = \{\mathbf{0}\}$, repeated application of the dimension formula (Corollary 2.3.19) gives

$$\begin{aligned} \dim \left(\sum_{i=1}^\ell \mathcal{K}_{d_i}(D, \mathbf{x}_i) \right) &= \dim(\mathcal{N}(D) + \mathcal{R}(D) + \text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k)) \\ &= \dim(\mathcal{N}(D) + \mathcal{R}(D)) + \dim(\text{span}(\mathbf{x}_1, \dots, \mathbf{x}_k)) \\ &= \dim(\mathcal{N}(D)) + \dim(\mathcal{R}(D)) + k \\ &= n - \dim(\mathcal{N}(D) \cap \mathcal{R}(D)) + k \\ &= n. \end{aligned}$$

Therefore

$$\bigoplus_{i=1}^\ell \mathcal{K}_{d_i}(D, \mathbf{x}_i) = V. \quad \square$$

12.10.2 Jordan Normal Form

Combining the two previous propositions with the spectral decomposition yields an almost-diagonal form for any finite-dimensional operator.

Theorem 12.10.4 (Jordan Normal Form). *Let V be a finite-dimensional vector space, and let A be an operator on V with spectral decomposition*

$$A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda.$$

For each $\lambda \in \sigma(A)$ let $\mathcal{E}_\lambda = \mathcal{R}(P_\lambda)$ be the generalized eigenspace of A associated with λ and let $m_\lambda = \text{ind}(D_\lambda)$. For $d_1, \dots, d_\ell$ as in the previous proposition, there is a basis for $\mathcal{E}_\lambda$ in which $\lambda P_\lambda + D_\lambda$ is represented by an $m_\lambda \times m_\lambda$ matrix that can be written in block-matrix form as

$$M_\lambda = \begin{bmatrix} J_{d_1} & 0 & \cdots & 0 \\ 0 & J_{d_2} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & J_{d_\ell} \end{bmatrix}$$

and each J_{d_i} is a $d_i \times d_i$ matrix of the form

$$J_{m_i} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}.$$

In particular, if $A \in M_n(\mathbb{F})$, then there exists an invertible matrix $S \in M_n(\mathbb{C})$ such that $A = S^{-1}MS$, where M has the block-diagonal form

$$M = \begin{bmatrix} M_{\lambda_1} & 0 & \cdots & 0 \\ 0 & M_{\lambda_2} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & M_{\lambda_r} \end{bmatrix},$$

where r is the number of distinct eigenvalues in $\sigma(A)$.

Remark 12.10.5. The Jordan form is poorly suited to computation because the problem of finding the Jordan form is ill conditioned. This is because bases of the form $\mathbf{x}, D\mathbf{x}, D^2\mathbf{x}, \dots$ are usually far from being orthogonal, with many almost-parallel vectors. We show in section 13.3 that an orthogonal basis, such as the Arnoldi basis, gives much better computational results. Of course the matrix forms arising from Arnoldi bases do not have the nice visual appeal of the Jordan form, but they are generally much more useful.

Remark 12.10.6 (Finding the Jordan normal form). Here we summarize the algorithm for finding the Jordan form for a given matrix $A \in M_n(\mathbb{F})$.

- (i) Find the spectrum $\{\lambda_1, \dots, \lambda_r\}$ of A and the algebraic multiplicity m_j of each λ_j . (When working by hand, on very small matrices, this could be done by computing the characteristic polynomial $p_A(z) = \det(zI - A)$ and factoring the polynomial as $\prod_{j=1}^r (z - \lambda_j)^{m_j}$, but for most matrices, other algorithms, such as those of Section 13.4, are both faster and more stable.)
- (ii) For each λ_j , compute the dimension of $\mathcal{N}((\lambda_j I - A)^k)$ for $k = 1, 2, \dots$ until $\dim \mathcal{N}((\lambda_j I - A)^k) = m_j$. Setting $d_{j_k} = \dim \mathcal{N}((\lambda_j I - A)^k)$, this gives a sequence $0 = d_{j_0} < d_{j_1} < \dots < d_{j_r} = m_j$.
- (iii) The sequence (d_{j_r}) gives information about the sizes of the corresponding blocks. The sequence is interpreted as follows: $d_{j_1} - d_{j_0} = d_{j_1}$ tells how many Jordan blocks of size at least one corresponding to eigenvalue λ_j there are. Similarly $d_{j_2} - d_{j_1}$ tells how many Jordan blocks are at least size 2, and so forth.
- (iv) Construct the matrix J using the information from step (iii).

Example 12.10.7. Let

$$A = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 3 & 3 \\ -1 & -2 & -2 \end{bmatrix}.$$

Applying the algorithm above, we have the following:

- (i) The characteristic polynomial of A is $p_A(z) = (z - 1)^3$; hence A has a single eigenvalue $\lambda = 1$ with multiplicity 3.
- (ii) Since $\lambda = 1$ we have

$$1I - A = \begin{bmatrix} -1 & -2 & -3 \\ -1 & -2 & -3 \\ 1 & 2 & 3 \end{bmatrix}$$

and $d_1 = \dim \mathcal{N}((1I - A)^1) = 2$. Since $(1I - A)^2 = 0$, we have $d_2 = \dim \mathcal{N}((1I - A)^2) = 3$, which is equal to the multiplicity of λ .

- (iii) Since $d_0 = 0$, $d_1 = 2$, and $d_2 = 3$, there are at least two Jordan blocks with size at least one, and one Jordan block with size at least two.

Hence the Jordan normal form of A is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Example 12.10.8. Let A be the matrix in Example 12.2.4, that is,

$$A = \begin{bmatrix} 0 & -1 & 1 & 1 \\ 3 & -2 & -4 & 5 \\ 1 & -1 & -1 & 2 \\ 2 & -1 & -3 & 3 \end{bmatrix}.$$

To find the Jordan normal form, we first verify that $\sigma(A) = \{0\}$. We already saw that $A^2 = 0$, and that $\dim \mathcal{N}(A) = 2$, so the geometric multiplicity of the eigenvalue $\lambda = 0$ is 2. This shows that $\mathcal{E}_0$ can be decomposed into two complementary two-dimensional subspaces. The 0-eigenspace is $\mathcal{N}(A) = \text{span}\{\mathbf{x}_1, \mathbf{x}_3\}$, spanned by $\mathbf{x}_1 = [2 \ 1 \ 1 \ 0]^T$ and $\mathbf{x}_3 = [-1 \ 1 \ 0 \ 1]^T$.

We seek an $\mathbf{x}_2$ so that $\{(A - \lambda I)\mathbf{x}_2, \mathbf{x}_2\} = \{\mathbf{x}_1, \mathbf{x}_2\}$ and an $\mathbf{x}_4$ so that $\{(A - \lambda I)\mathbf{x}_4, \mathbf{x}_4\} = \{\mathbf{x}_3, \mathbf{x}_4\}$. So we must solve $(A - 0I)\mathbf{x}_2 = A\mathbf{x}_2 = \mathbf{x}_1$ and $A\mathbf{x}_4 = \mathbf{x}_3$. We find $\mathbf{x}_2 = [-1 \ -2 \ 0 \ 0]^T$ and $\mathbf{x}_4 = [1 \ 1 \ 0 \ 0]^T$. Verify that the set $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \{A\mathbf{x}_2, \mathbf{x}_2, A\mathbf{x}_4, \mathbf{x}_4\}$ is a basis and that this basis gives the desired normal form; that is, verify that we have

$$J = S^{-1}AS = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{where } S = \begin{bmatrix} 2 & -1 & -1 & 1 \\ 1 & -2 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The verification of the preceding details is Exercise 12.51.

Example 12.10.9. Consider the matrix

$$A = \begin{bmatrix} -1 & 0 & -4 \\ -2 & 1 & -4 \\ 1 & 0 & 3 \end{bmatrix}.$$

First verify that $\sigma(A) = \{1\}$ and $\mathcal{N}(I - A) = \text{span}\{\mathbf{z}_1, \mathbf{z}_2\}$, where $\mathbf{z}_1 = [0 \ 1 \ 0]^T$ and $\mathbf{z}_2 = [-2 \ 0 \ 1]^T$.

Since $(I - A)^2 = 0$, we have two blocks that are $(I - A)$ -invariant: a 1×1 block and a (not semisimple) 2×2 block. To find the 2×2 block we must find an $\mathbf{x}$ that is not in $\mathcal{N}((A - \lambda I))$. This will guarantee that $\{(A - \lambda I)\mathbf{x}, \mathbf{x}\}$ has two elements. In our case a simple choice is $\mathbf{x}_2 = [1 \ 0 \ 0]^T$. Let $(A - \lambda I)\mathbf{x}_2 = \mathbf{x}_1$. Since $(A - I)^2 = 0$, we have $(A - I)\mathbf{x}_1 = \mathbf{0}$, which gives $A\mathbf{x}_1 = \mathbf{x}_1$ and $A\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{x}_2$.

Verify that $\mathbf{z}_1$ is not in $\text{span}(\mathbf{x}_1, \mathbf{x}_2)$, so $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1\}$ gives a basis of $\mathbb{F}^3$. Verify that letting $P = [\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{z}_1]$ puts $P^{-1}AP$ in Jordan form. The verification of the preceding details is Exercise 12.52.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

12.1. Verify the claim of Example 12.1.2; that is, show that $(I - P)^2 = I - P$.

12.2. Show that the matrix

$$\frac{1}{2} \begin{bmatrix} 2 \cos^2 \theta & \sin 2\theta \\ \sin 2\theta & 2 \sin^2 \theta \end{bmatrix}$$

is a projection for all $\theta \in \mathbb{R}$.

12.3. Let P be a projection on a finite-dimensional vector space V . Prove that $\text{rank}(P) = \text{tr}(P)$. Hint: Use Corollary 12.1.5 and Exercise 2.25.

12.4. Let $F \in M_n(\mathbb{F})$ be the $n \times n$ matrix that reverses the indexing of $\mathbf{x}$, that is, if $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^\top$, then $F\mathbf{x} = [x_n \ x_{n-1} \ \dots \ x_1]^\top$. Define $E \in M_n(\mathbb{F})$ such that $E\mathbf{x} = (\mathbf{x} + F\mathbf{x})/2$. Prove that E is a projection. Determine whether it is an orthogonal projection or an oblique projection. What are the entries of E ?

12.5. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 5 & 7 \end{bmatrix}.$$

Find the eigenvalues and eigenvectors of A and use those to write out the eigenprojection matrices P_1 and P_2 of Proposition 12.1.10. Verify that each of the five properties listed in that proposition holds for these matrices.

12.6. Let L be a linear operator on a finite-dimensional vector space V . Prove that $\text{ind}(L) \leq \dim(V)$.

12.7. Assume $A \in M_n(\mathbb{R})$ is positive semidefinite. Show that if $\mathbf{x}$ is a unit vector, then $\mathbf{x}^\top A \mathbf{x}$ is less than or equal to the largest eigenvalue. Hint: Write $\mathbf{x}$ as a linear combination of an orthonormal eigenbasis of A .

12.8. Find a basis of generalized eigenvectors of

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ 0 & -1 & 1 \end{bmatrix}.$$

12.9. Assume that λ and μ are eigenvalues of $A \in M_n(\mathbb{F})$. If $A\mathbf{x} = \mu\mathbf{x}$ and $(\lambda I - A)^m \mathbf{x} = \mathbf{0}$, then $(\lambda - \mu)^m \mathbf{x} = \mathbf{0}$. Hint: Use the binomial theorem.

12.10. Recall from Theorem 12.2.5 that if $k = \text{ind}(B)$, then $\mathbb{F}^n = \mathcal{R}(B^k) \oplus \mathcal{N}(B^k)$. Prove that the two subspaces $\mathcal{R}(B^k)$ and $\mathcal{N}(B^k)$ are B -invariant.

12.11. Assume $N \in M_n(\mathbb{F})$ is nilpotent (see Exercise 4.1), that is, $N^k = 0$ for some $k \in \mathbb{N}$.

- (i) Use the Neumann series of $I + N$ to prove that $I + N$ is invertible.
- (ii) Write $(I + N)^{-1}$ in terms of powers of N .
- (iii) Generalize your results to show that $\lambda I + N$ is invertible for any nonzero $\lambda \in \mathbb{C}$.
- (iv) Write $(\lambda I + N)^{-1}$ in terms of powers of N .

12.12. Let

$$A = \begin{bmatrix} 0 & 2 \\ 6 & 1 \end{bmatrix}.$$

- (i) Write the resolvent $R(z)$ of A in the form $R(z) = B(z)/p(z)$, where $B(z)$ is a matrix of polynomials in z and $p(z)$ is a monic polynomial in z .
- (ii) Factor $p(z)$ into the form $p(z) = (z - a)(z - b)$ and write the partial fraction decomposition of $R(z)$ in the form $R(z) = C/(z - a) + D/(z - b)$, where C and D are constant matrices.

12.13. Let

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 7 \end{bmatrix}.$$

- (i) Write the resolvent $R(z)$ of A in the form $R(z) = B(z)/p(z)$, where $B(z)$ is a matrix of polynomials in z and $p(z)$ is a monic polynomial in z .
- (ii) Factor $p(z)$ and write the partial fraction decomposition of $R(z)$.

12.14. Let

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 7 \end{bmatrix}.$$

Compute the spectral radius $r(A)$ using the 2-norm (see Exercise 4.16), using the 1-norm $\|\cdot\|_1$ (see Theorem 3.5.20), and using the infinity norm $\|\cdot\|_\infty$.

12.15. Let $\|\cdot\|$ be a matrix norm. Prove that if $A \in M_n(\mathbb{F})$ satisfies $r(A) < 1$, then for any $0 < \varepsilon < 1 - r(A)$ there exists $m \in \mathbb{Z}^+$ such that

$$\|A^k\| < (r(A) + \varepsilon)^k \leq r(A) + \varepsilon < 1$$

for all $k \geq N$.

12.16. For $A \in M_n(\mathbb{F})$, let $|A|$ denote componentwise modulus, so if a_{ij} is the (i, j) entry of A , then $|a_{ij}|$ is the (i, j) entry of $|A|$. Let $\succeq$ denote componentwise inequality (see also Definition 12.8.1).

(i) Prove that $r(A) \leq r(|A|)$.

(ii) Prove that if $0 \preceq A \preceq B$, then $r(A) \leq r(B)$.

12.17. Using the matrix A in Exercise 12.12, compute its spectral projections and show that the four properties in Theorem 12.4.2 are satisfied.

12.18. Using the matrix A in Exercise 12.13, compute its spectral projections and show that the four properties in Theorem 12.4.2 are satisfied.

12.19. Verify the Cayley–Hamilton theorem for the matrix A in Exercise 12.12.

12.20. Compute the eigenvalues, resolvent, and spectral projections of the matrix

$$B = \begin{bmatrix} 2 & 4 & 6 & 8 \\ 0 & 2 & 4 & 6 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

12.21. Let

$$A = \begin{bmatrix} 0 & 3i \\ 3i & 6 \end{bmatrix}.$$

Find the Laurent expansion of the resolvent of A around the point $z = 3$.

12.22. Let B be the matrix in Exercise 12.20.

(i) Find the Laurent expansion of the resolvent of B around the point $z = 2$.

(ii) Find the Laurent expansion of the resolvent of B around the point $z = 4$.

12.23. Prove Lemma 12.5.5, parts (i) and (ii).

12.24. Prove Lemma 12.5.5, part (iii).

12.25. Prove Lemma 12.5.5, parts (iv) and (v).

12.26. Let A be the matrix in Exercise 12.12.

(i) Write the spectral decomposition $A = \sum_{\lambda} \lambda P_{\lambda} + D_{\lambda}$.

(ii) Write the resolvent in the form of (12.33).

(iii) Write a general formula for A^k in a form that involves no matrix multiplications.

12.27. Let B be the matrix in Exercise 12.20.

(i) Write the spectral decomposition $B = \sum_{\lambda} \lambda P_{\lambda} + D_{\lambda}$.

(ii) Write the resolvent in the form of (12.33).

(iii) Compute $\cos(B\pi/6)$.

12.28. Prove that the number of distinct n th roots of an invertible matrix A is at least n^k , where k is the number of distinct eigenvalues of A .

12.29. Let $A \in M_n(\mathbb{F})$. Prove that if $A\mathbf{v} = \lambda\mathbf{v}$, then $\mathbf{v} \in \mathcal{N}(D_{\lambda})$.

- 12.30. Given $A \in M_n(\mathbb{F})$, use the spectral decomposition (12.33) and Exercise 12.3 to prove that $\operatorname{tr} R(z) = p'(z)/p(z)$, where p is the characteristic polynomial of A . Hint: Write the characteristic polynomial in factored form (4.5), and show that

$$\frac{p'(z)}{p(z)} = \sum_{i=1}^r \frac{m_i}{z - \lambda_i}.$$

-
- 12.31. Compute e^A for each A in Exercises 12.12 and 12.13.
- 12.32. Let $\mathbf{x} \in \mathbb{F}^n$ be an eigenvector of A associated to the eigenvalue $\lambda \in \sigma(A)$. Let $U \subset \mathbb{C}$ be an open, simply connected set containing $\sigma(A)$, and let $f(z)$ be a holomorphic function on U . Show that $\mathbf{x}$ is an eigenvector of $f(A)$ with eigenvalue $f(\lambda)$. Under what circumstances could one have an eigenvector $\mathbf{y}$ of $f(A)$ with eigenvalue $f(\lambda)$ where $\mathbf{y}$ is *not* an eigenvector of A ?
- 12.33. Prove that if $U \subset \mathbb{C}$ is an open, simply connected set containing $\sigma(A)$, and if $f(z)$ is a holomorphic function on U such that $f(A) = 0$, then $f(\lambda) = 0$ for every $\lambda \in \sigma(A)$. Then prove that the converse is false.
- 12.34. Let $B \in M_n(\mathbb{F})$. Assume $\lambda \in \sigma(B)$ is a semisimple eigenvalue of modulus $|\lambda| = r(B)$, with corresponding eigenprojection P , and that all other eigenvalues have modulus strictly less than $r(B)$. Let $\mathbf{x}_0$ be a unit vector satisfying $P\mathbf{x}_0 \neq \mathbf{0}$. For each $k \in \mathbb{N}$, define

$$\mathbf{x}_{k+1} = \frac{B\mathbf{x}_k}{\|B\mathbf{x}_k\|}. \quad (12.61)$$

Prove that $(\mathbf{x}_k)_{k=0}^\infty$ has a convergent subsequence, that the corresponding limit $\mathbf{x}$ is an eigenvector of B corresponding to λ , and that $\lambda = \mathbf{x}^H B \mathbf{x}$. Hint: The matrix $A = \lambda^{-1}B$ satisfies the hypotheses in Theorem 12.7.8 and has the same eigenprojections as B . Consider using the fact that $\|A^k \mathbf{x}_0 - P\mathbf{x}_0\| \rightarrow 0$ to show that

$$\left\| \mathbf{x}_k - \frac{P\mathbf{x}_0}{\|P\mathbf{x}_0\|} \right\| \rightarrow 0.$$

- 12.35. Give an example of a matrix $A \in M_2(\mathbb{F})$ that has two distinct eigenvalues of modulus 1 and such that the iterative map $\mathbf{x}_k = A\mathbf{x}_{k-1}$ fails to provide an eigenvector for any nonzero $\mathbf{x}_0$ that is not already an eigenvector.
-

- 12.36. Let

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 3 & 0 & 3 \\ 0 & 2 & 0 \end{bmatrix}.$$

- (i) Show that A is irreducible.
 (ii) Find the Perron root of A .

- 12.37. For $A \in M_n(\mathbb{F})$, prove that $A \succ 0$ if and only if $A\mathbf{x} \succ \mathbf{0}$ for every $\mathbf{x} \succeq \mathbf{0}$ with $\mathbf{x} \neq \mathbf{0}$.
- 12.38. Prove Proposition 12.8.10.

12.39. Assume $A \succeq 0$, $A \in M_n(\mathbb{F})$.

- (i) Prove that if A is primitive, then $r(A)$ is an eigenvalue of A and that no other eigenvalue has norm $r(A)$.
- (ii) Prove that if $A \succeq 0$ is irreducible, then its Perron root is simple, and the corresponding eigenvector $\mathbf{x}$ is positive ($\mathbf{x} \succ \mathbf{0}$).
- (iii) Give an example showing that, unlike the case of positive matrices, there are irreducible matrices with more than one eigenvalue lying on the circle $|z| = r(A)$.

12.40. Let $A \in M_n(\mathbb{F})$. Prove that if $A \succeq 0$ is irreducible with Perron root $\lambda = r(A)$ and corresponding eigenvector $\mathbf{x} \succ \mathbf{0}$, then for any other nonnegative real eigenvector $\mathbf{y} \succeq \mathbf{0}$, the corresponding eigenvalue μ must be equal to $r(A)$ and $\mathbf{y} = c\mathbf{x}$ for some positive real constant c . To prove this, follow these steps:

- (i) Show that if $A \succeq 0$ is irreducible, then $A^T \succeq 0$ is also irreducible, and $r(A) = r(A^T)$.
- (ii) Show that if $\mathbf{z}$ is the positive eigenvector of A^T associated to the eigenvalue $r(A^T)$, then $\mathbf{z}^T \mathbf{y} > 0$.
- (iii) Show that $r(A)\mathbf{z}^T \mathbf{y} = \mathbf{z}^T A\mathbf{y} = \mu \mathbf{z}^T \mathbf{y}$, and hence $\mu = r(A)$.
- (iv) Prove that $\mathbf{y} \in \text{span}(\mathbf{x})$.

12.41. Use the previous exercise to show that if the row sums $\sum_j a_{ij}$ of an irreducible matrix $A = (a_{ij}) \succeq 0$ are all the same for every i , then for every i we have $r(A) = \sum_j a_{ij}$.

12.42. Let $B \in M_n(\mathbb{R})$ be a positive matrix where every column's sum equals one. Prove: If P is the eigenprojection of B corresponding to the Perron root $\lambda = 1$, then $P\mathbf{v} \neq \mathbf{0}$ for any nonzero, nonnegative $\mathbf{v} \in \mathbb{R}^n$. Hint: Write the eigenprojection in terms of the right and left eigenvectors, noting that the left eigenvector is $\mathbf{1}$.

12.43. Given the matrix A from Example 12.9.6, use the Wedderburn decomposition to compute A^D . In particular, let

$$S = \begin{bmatrix} 1 & 0 & 0 & -27 \\ 0 & 1 & 0 & 9 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad S^{-1} = \begin{bmatrix} 1 & 0 & 0 & 27 \\ 0 & 1 & 0 & -9 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Verify that your answer is the same as that given in Example 12.9.6.

12.44. Given $A \in M_n(\mathbb{F})$, let P_0 and D_0 denote, respectively, the eigenprojection and the eigennilpotent of the eigenvalue $\lambda = 0$. Show that

$$A^D = (A - D_0 + P_0)^{-1} - P_0.$$

12.45. Given $A \in M_n(\mathbb{F})$, let P_0 and D_0 denote, respectively, the eigenprojection and the eigennilpotent of the eigenvalue $\lambda = 0$. Also let $P_* = I - P_0$. Prove that the following properties of the Drazin inverse hold:

- (i) $P_0 A^D = A^D P_0 = 0$.

- (ii) $D_0 A^D = A^D D_0 = 0$.
 (iii) $P_* A^D = A^D P_* = A^D$.
 (iv) $A^D (A - D_0) = P_* = (A - D_0) A^D$.
- 12.46. Given $A \in M_n(\mathbb{F})$, prove that the following properties of the Drazin inverse hold:
- (i) $(A^D)^H = (A^H)^D$.
 (ii) $(A^D)^\ell = (A^\ell)^D$ for any $\ell \in \mathbb{N}$.
 (iii) $A^D = A$ if and only if $A^3 = A$.
 (iv) $(A^D)^D = A^2 A^D$.
- 12.47. Given $A \in M_n(\mathbb{F})$, prove that the following properties of the Drazin inverse hold:
- (i) $A^D = 0$ if and only if A is nilpotent.
 (ii) $\text{ind}(A^D) = 0$ if and only if A is nonsingular.
 (iii) $\text{ind}(A^D) = 1$ if and only if A is singular.
 (iv) If A is idempotent, then $A^D = A$.
- 12.48. Let $A \in M_n(\mathbb{F})$ and $m = \text{ind}(A)$. Prove the following:
- (i) $P_* = A^D A$ is the projection onto $\mathcal{R}(A^m)$ along $\mathcal{N}(A^m)$. Hint: Show that $A^m = P_* A^m = A^m P_*$ and use the fact that A is invertible on $\mathcal{R}(P_*)$.
 (ii) If $\mathbf{b} \in \mathcal{R}(A^m)$, then $\mathbf{x} = A^D \mathbf{b}$ is a solution of $A\mathbf{x} = \mathbf{b}$.
- 12.49. Prove that if $\mathbf{b} \in \mathcal{R}(P_*)$, then the linear system $A\mathbf{x} = \mathbf{b}$ has the solution $\mathbf{x} = A^D \mathbf{b}$.

12.50.* Prove Proposition 12.10.2.

12.51.* Fill in all the details for Example 12.10.8.

12.52.* Fill in all the details for Example 12.10.9.

12.53.* Given the matrix

$$A = \begin{bmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 2 \\ 0 & 0 & \alpha \end{bmatrix},$$

find the matrix P such that

$$P^{-1}AP = \begin{bmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{bmatrix}.$$

Notes

Two great sources on resolvent methods are [Kat95] and [Cha12]. Exercise 12.4 comes from [TB97].

A short history and further discussion of Perron's theorem and the Perron–Frobenius theorem can be found in [Mac00]. For more details about Google's PageRank algorithm, see [LM06] and [BL06].

For more about the Drazin inverse, see [CM09] and [Wil79]. The proof of Proposition 12.9.10 is modeled after the proof of the same result in [Wil79].

The proof of Proposition 12.10.3 was inspired by [Wil07].

13 Iterative Methods

Have no fear of perfection—you'll never reach it.
—Salvador Dali

The subject of numerical linear algebra is concerned with computing approximate answers to linear algebra problems. The two main numerical linear algebra problems are solving linear systems and finding eigenvalues and eigenvectors.

For small matrices, naïve methods, such as Gaussian elimination, are generally suitable because computers can solve these little problems in the blink of an eye. However, for large matrices, naïve methods can take a lifetime to give a solution. For example, the temporal complexity (number of basic computations) of Gaussian elimination is cubic, meaning that if the dimension of the matrix doubles, then the run time increases by a factor of eight. It doesn't take very many of these doublings before the run time is too long and we start to wish for algorithms with lower temporal complexity.

In addition to the issue of temporal complexity, the spatial complexity—the amount of memory, or storage space, required to solve the problem—is also important. The spatial complexity of matrix storage is quadratic, meaning that as the matrix dimension doubles, the amount of space required to store the matrix increases by a factor of four. It's very possible to encounter linear algebra problems that are so large that storing the entire matrix on a single computer is impossible, and solvers must therefore distribute the workload across multiple machines.

Fortunately, the matrices considered in most real-world applications are *sparse*, that is, they have relatively few nonzero entries. Methods that take advantage of this sparsity greatly improve the run times and do so with much less memory than methods that don't. These sparse methods store only the nonzero entries and therefore use only a tiny fraction of the memory that would be needed to store all the entries. Moreover, the matrix operations used in row reduction or matrix multiplication can skip over the zeros since they don't alter the computation; when most of the entries are zero, that skipping speeds up the algorithms substantially.

Numerical methods for large sparse linear algebra problems tend to be *iterative*. These algorithms work by generating a sequence of successive approximations that converge to the true solution, where each iteration builds on the previous

approximation, until it is sufficiently close that the algorithm can terminate. For example, Newton's method, described in Section 7.3, is an iterative method for finding zeros.

We begin the chapter with a result about convergence of sequences of the form

$$\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}, \quad k \in \mathbb{N},$$

for $B \in M_n(\mathbb{F})$ and $\mathbf{x}_0, \mathbf{c} \in \mathbb{F}^n$. This one result gives three fundamental iterative methods for solving linear systems: the Jacobi method, the Gauss–Seidel method, and the method of successive overrelaxation (SOR).

We then develop some iterative methods for solving numerical linear algebra problems based on *Krylov subspaces*. Krylov methods rank among the fastest general methods available for solving large, sparse linear systems and eigenvalue problems. One of the most important of these methods is the *GMRES algorithm*.

We conclude the chapter with some powerful iterative methods for computing eigenvalues, including QR iteration and the Arnoldi method.

13.1 Methods for Linear Systems

We have already seen an example of an iterative method with the power method (see Section 12.7.3) for finding the Perron eigenvector. This eigenvector is found by computing successive powers $\mathbf{x}_0, B\mathbf{x}_0, B^2\mathbf{x}_0, \dots$, for any $\mathbf{x}_0 \succeq \mathbf{0}$ whose entries sum to one. We can think of this as a sequence $(\mathbf{x}_k)_{k=0}^\infty$ generated by the rule $\mathbf{x}_{k+1} = B\mathbf{x}_k$ for each $k \in \mathbb{N}$. The limit $\mathbf{x}$ of this sequence is a fixed point of B ; that is, $\mathbf{x} = B\mathbf{x}$.

It is useful to generalize this result to sequences that are generated by successive powers together with a shift, that is,

$$\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}, \quad k \in \mathbb{N}, \tag{13.1}$$

where $B \in M_n(\mathbb{F})$ and $\mathbf{c}, \mathbf{x}_0 \in \mathbb{F}^n$ are given. Iterative processes of this form show up in applications such as Markov chains, analysis of equilibria in economics,⁴⁸ and control theory.

In this section we first show that sequences of the form (13.1) converge whenever the spectral radius $r(B)$ satisfies $r(B) < 1$, regardless of the initial starting point, and that the limit $\mathbf{x} \in \mathbb{F}^n$ satisfies the equation $\mathbf{x} = B\mathbf{x} + \mathbf{c}$, or, equivalently, $\mathbf{x} = (I - B)^{-1}\mathbf{c}$. We then use this fact to describe three iterative methods for approximating solutions to linear systems of the form $A\mathbf{x} = \mathbf{b}$.

13.1.1 Convergence Theorem

For any $B \in M_n(\mathbb{F})$, if $\|B\| < 1$, then Exercise 7.9 shows that the map $\mathbf{x} \mapsto B\mathbf{x} + \mathbf{c}$ is a contraction mapping on $\mathbb{F}^n$ and the corresponding fixed point is $(I - B)^{-1}\mathbf{c}$, provided $\|\cdot\|$ is the norm induced on $M_n(\mathbb{F})$ by the norm on $\mathbb{F}^n$. Moreover, the method of successive approximations (see Remark 7.1.10) always converges to the fixed point of a contraction mapping, and thus sequences generated by (13.1)

⁴⁸In 1973, Wassily Leontief won the Nobel Prize in economics for his work on input-output analysis, which used successive powers to compute equilibria for the global production of commodities.

converge to $\mathbf{x} = (I - B)^{-1}\mathbf{c}$ whenever $\|B\| < 1$. The next theorem shows that this convergence result holds whenever $r(B) < 1$, regardless of the size of $\|B\|$.

Theorem 13.1.1. *Let $B \in M_n(\mathbb{F})$ and $\mathbf{c} \in \mathbb{F}^n$. If $r(B) < 1$, then sequences generated by (13.1) converge to $\mathbf{x} = (I - B)^{-1}\mathbf{c}$, regardless of the initial vector $\mathbf{x}_0$.*

Proof. Let $\|\cdot\|$ be the induced norm on $M_n(\mathbb{F})$. If $r(B) < 1$, then Exercise 12.15 shows that for any $0 < \varepsilon < 1 - r(B)$, there exists $m \in \mathbb{Z}^+$ such that $\|B^k\| < (r(B) + \varepsilon)^k \leq r(B) + \varepsilon < 1$ for all $k \geq m$. Set $f(\mathbf{x}) = B\mathbf{x} + \mathbf{c}$ and observe that the k th composition satisfies

$$f^k(\mathbf{x}) = B^k\mathbf{x} + \sum_{j=0}^{k-1} B^j\mathbf{c}.$$

Thus, for any $k \geq m$, we have

$$\|f^k(\mathbf{x}) - f^k(\mathbf{y})\| = \|B^k(\mathbf{x} - \mathbf{y})\| \leq \|B^k\|\|\mathbf{x} - \mathbf{y}\| \leq (r(B) + \varepsilon)\|\mathbf{x} - \mathbf{y}\|,$$

and therefore f^k is a contraction mapping. By Theorem 7.1.14 and the solution of Exercise 7.6, the sequence $(\mathbf{x}_k)_{k=0}^\infty$ converges to the unique fixed point $\mathbf{x}$ of f , and thus $\mathbf{x}$ satisfies $\mathbf{x} = B\mathbf{x} + \mathbf{c}$. Since $I - B$ is invertible (otherwise $\lambda = 1$ is an eigenvalue of B , which would contradict $r(B) < 1$), we have $\mathbf{x} = (I - B)^{-1}\mathbf{c}$. $\square$

Remark 13.1.2. The proof of the previous theorem also shows that the sequence (13.1) converges linearly (see Definition 7.3.1), and it gives a bound on the approximation error $\varepsilon_k = \|\mathbf{x}_k - \mathbf{x}_0\|$, namely,

$$\varepsilon_k = \|\mathbf{x} - \mathbf{x}_k\| = \|f^k(\mathbf{x}) - f^k(\mathbf{x}_0)\| = \|B^k(\mathbf{x} - \mathbf{x}_0)\| \leq \|B^k\|\|\mathbf{x} - \mathbf{x}_0\| \leq (r(B) + \varepsilon)^k \varepsilon_0.$$

13.1.2 Iterative Solvers

Theorem 13.1.1 lies at the heart of three different iterative solvers for linear systems, namely, the *Jacobi method*, the *Gauss-Seidel method*, and the method of *successive overrelaxation* (SOR). These all follow from the same idea. First, decompose a linear system $A\mathbf{x} = \mathbf{b}$ into an expression of the form $\mathbf{x} = B\mathbf{x} + \mathbf{c}$, where $r(B) < 1$. Then, by Theorem 13.1.1, the iteration (13.1) converges to the solution linearly.

Throughout this section, assume that $A \in M_n(\mathbb{F})$ is given and that A is decomposed into a diagonal part D , a strictly upper-triangular part U , and a strictly lower-triangular part L , so that $A = L + D + U$.

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ * & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & 0 \end{bmatrix} + \begin{bmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{bmatrix} + \begin{bmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Here $*$ indicates an arbitrary entry—not necessarily zero or nonzero.

Jacobi's Method

By writing $A\mathbf{x} = \mathbf{b}$ as $D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}$, we have

$$\mathbf{x} = -D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}, \quad (13.2)$$

which motivates the definition $B_{\text{Jac}} = -D^{-1}(L + U)$ and $\mathbf{c}_{\text{Jac}} = D^{-1}\mathbf{b}$. Jacobi's method is to apply the iteration (13.1) with B_{Jac} and $\mathbf{c}_{\text{Jac}}$. Note that D^{-1} is trivial to compute because D is diagonal, and multiplication by D^{-1} is also trivial. Theorem 13.1.1 guarantees that Jacobi's method converges to $\mathbf{x}$ as long as $r(B_{\text{Jac}}) < 1$.

Example 13.1.3. Consider the linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}.$$

Decomposing A , we have

$$D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad U = \begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus,

$$B_{\text{Jac}} = -D^{-1}(L + U) = \frac{-1}{6} \begin{bmatrix} 0 & 4 & 2 \\ 6 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{c}_{\text{Jac}} = D^{-1}\mathbf{b} = \frac{1}{2} \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix}.$$

Starting with $\mathbf{x}_0 = [0 \ 1 \ 1]^T$ and iterating via (13.1), we find that $\mathbf{x}_k$ approximates the correct answer $[-7 \ 9 \ 6]^T$ to four digits of accuracy once $k \geq 99$. We note that $r(B_{\text{Jac}}) = \sqrt{5/6} \approx 0.9129$, which implies that the approximation error satisfies $\|\mathbf{x} - \mathbf{x}_k\| \leq 0.9130^k \cdot \|\mathbf{x} - \mathbf{x}_0\|$ for k sufficiently large.

The Gauss–Seidel Method

By writing $A\mathbf{x} = \mathbf{b}$ as $(D + L)\mathbf{x} = -U\mathbf{x} + \mathbf{b}$, we have that

$$\mathbf{x} = -(D + L)^{-1}U\mathbf{x} + (D + L)^{-1}\mathbf{b}. \quad (13.3)$$

Hence for $B_{\text{GS}} = -(D + L)^{-1}U$ and $\mathbf{c} = (D + L)^{-1}\mathbf{b}$, we have that the iteration (13.1) converges to $\mathbf{x}$ as long as $r(B_{\text{GS}}) < 1$. Note that although the inverse $(D + L)^{-1}$ is fairly easy to compute (because $D + L$ is lower triangular), it is generally both faster and more stable to write the Gauss–Seidel iteration as

$$(D + L)\mathbf{x}_{k+1} = -U\mathbf{x}_k + \mathbf{b}$$

and then solve for $\mathbf{x}_{k+1}$ by back substitution.

Example 13.1.4. Considering the system from Example 13.1.3, we have

$$B_{\text{GS}} = -(D + L)^{-1}U = \frac{1}{6} \begin{bmatrix} 0 & -4 & -2 \\ 0 & 4 & 2 \\ 0 & 2 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{c}_{\text{GS}} = (D + L)^{-1}\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}.$$

Starting with $\mathbf{x}_0 = [0 \ 1 \ 1]^T$ and iterating via (13.1), we find that $\mathbf{x}_k$ approximates the correct answer $[-7 \ 9 \ 6]^T$ to four digits of accuracy once $k \geq 50$.

Here we have $r(B_{\text{GS}}) = 5/6 \approx 0.8333$, which is the square of the radius for the Jacobi method in the previous example. This implies that the approximation error satisfies $\|\mathbf{x} - \mathbf{x}_k\| \leq 0.9130^{2k} \cdot \|\mathbf{x} - \mathbf{x}_0\|$. Therefore, in this example at least, the Gauss–Seidel method can solve this problem in half the number of iterations as the Jacobi method. In other words, it converges twice as fast.

Nota Bene 13.1.5. Just because one method converges faster than another does not necessarily mean that it is a faster algorithm. One must also look at the computational cost of each iteration.

Successive Overrelaxation

The Gauss–Seidel method converges faster than the Jacobi method because the spectral radius of B_{GS} is smaller than the spectral radius of B_{Jac} . We can often improve convergence even more by splitting up A in a way that has a free parameter and then tuning the decomposition to further reduce the value of $r(B)$. More precisely, by writing $A\mathbf{x} = \mathbf{b}$ as $(D + \omega L)\mathbf{x} = ((1 - \omega)D - \omega U)\mathbf{x} + \omega\mathbf{b}$, where $\omega > 0$, we have

$$\mathbf{x} = (D + \omega L)^{-1}((1 - \omega)D - \omega U)\mathbf{x} + \omega(D + \omega L)^{-1}\mathbf{b}. \quad (13.4)$$

Hence for $B_\omega = (D + \omega L)^{-1}((1 - \omega)D - \omega U)$ and $\mathbf{c}_\omega = \omega(D + \omega L)^{-1}\mathbf{b}$, the iteration (13.1) converges to $\mathbf{x}$ as long as $r(B_\omega) < 1$. We note of course that when $\omega = 1$ we have the Gauss–Seidel case. Choosing ω to make $r(B_\omega)$ as small as possible will give the fastest convergence.

Again, we note that rather than multiply by the inverse matrix to construct B_ω , it is generally both faster and more stable to write the iteration as

$$(D + \omega L)\mathbf{x}_{k+1} = ((1 - \omega)D - \omega U)\mathbf{x}_k + \omega\mathbf{b}$$

and then solve for $\mathbf{x}_{k+1}$ by back substitution.

Example 13.1.6. Consider again the system from Examples 13.1.3 and 13.1.4. Since $\omega = 1$ corresponds to the Gauss–Seidel case, we already know that $r(B_1) = 5/6 \approx 0.8333 \dots$

For a large class of matrices, the optimal choice of ω is

$$\omega_* = \frac{2}{1 + \sqrt{1 - \beta^2}}, \quad (13.5)$$

where $\beta = r(B_{\text{Jac}})$; see [Ise09, Gre97] for details. In the case of this example, $r(B_{\text{Jac}}) = \sqrt{5/6}$, which gives $\omega_* = \frac{2}{5}(6 - \sqrt{6}) \approx 1.4202$ and $r(B_{\omega_*}) = \omega_* - 1 \approx 0.4202$. Therefore, the approximation error is bounded by $\|\mathbf{x} - \mathbf{x}_k\| \leq 0.4202^k \cdot \|\mathbf{x} - \mathbf{x}_0\|$, so this method converges much faster than the Gauss–Seidel method.

For example, starting with $\mathbf{x}_0 = [0 \ 1 \ 1]^\top$, the term $\mathbf{x}_k$ approximates the correct answer $[-7 \ 9 \ 6]^\top$ to four digits of accuracy once $k \geq 18$. Compare this to $k \geq 50$ for Gauss–Seidel and $k \geq 99$ for Jacobi.

13.1.3 More Convergence Theorems

Definition 13.1.7. A matrix $A \in M_n(\mathbb{F})$ with components $A = [a_{ij}]$ is strictly diagonally dominant if for each i we have

$$\sum_{j \neq i} |a_{ij}| < a_{ii}. \quad (13.6)$$

Theorem 13.1.8. Let $A \in M_n(\mathbb{F})$ have components $A = [a_{ij}]$. If A is strictly diagonally dominant, then $r(B_{\text{Jac}}) < 1$, that is, Jacobi iteration (13.2) converges for any initial vector $\mathbf{x}_0$.

Proof. Assuming that $A = [a_{ij}]$, we have that

$$\|B_{\text{Jac}}\|_\infty = \|D^{-1}(L + U)\|_\infty = \max_{1 \leq i \leq n} \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|}.$$

Hence if A is strictly diagonally dominant, then for each i we have

$$\sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

which implies that $\|B_{\text{Jac}}\|_\infty < 1$. The result follows from Lemma 12.3.13. $\square$

Theorem 13.1.9. Let $A \in M_n(\mathbb{F})$ have components $A = [a_{ij}]$. If A is strictly diagonally dominant, then $r(B_{\text{GS}}) < 1$, and thus Gauss–Seidel iteration (13.3) converges for any initial vector $\mathbf{x}_0$.

Proof. Define

$$\gamma = \max_{1 \leq i \leq n} \frac{\sum_{j > i} |a_{ij}|}{|a_{ii}| - \sum_{j < i} |a_{ij}|}. \quad (13.7)$$

Since A is strictly diagonally dominant, we have that

$$\sum_{j>i} |a_{ij}| < |a_{ii}| - \sum_{j<i} |a_{ij}|,$$

which implies $\gamma < 1$. It suffices to show that $\|B_{GS}\|_\infty \leq \gamma$.

For any nonzero $\mathbf{x} \in \mathbb{F}^n$, set $\mathbf{y} = B_{GS}\mathbf{x}$. Thus, $(D + L)\mathbf{y} = -U\mathbf{x}$, or, equivalently,

$$a_{ii}y_i + \sum_{j<i} a_{ij}y_j = -\sum_{j>i} a_{ij}x_j$$

for every i . Choosing i such that $|y_i| = \|\mathbf{y}\|_\infty$, we have

$$\left(|a_{ii}| - \sum_{j<i} |a_{ij}| \right) \|\mathbf{y}\|_\infty \leq |a_{ii}| \|\mathbf{y}\|_\infty - \sum_{j<i} |a_{ij}| |y_j| \leq \left(\sum_{j>i} |a_{ij}| \right) \|\mathbf{x}\|_\infty,$$

which implies

$$\frac{\|B_{GS}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{\|\mathbf{y}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \frac{\sum_{j>i} |a_{ij}|}{|a_{ii}| - \sum_{j<i} |a_{ij}|} \leq \gamma.$$

Since this holds for all $\mathbf{x} \neq \mathbf{0}$, we have that $\|B_{GS}\|_\infty \leq \gamma < 1$. $\square$

Theorem 13.1.10. *If A is Hermitian and positive definite (denoted $A > 0$), then $r(B_\omega) < 1$ for all $\omega \in (0, 2)$. In other words, SOR converges for any initial vector $\mathbf{x}_0$.*

Proof. Since A is Hermitian, we have that $A = D + L + L^H$. It follows that

$$\begin{aligned} B_\omega &= (D + \omega L)^{-1}((1 - \omega)D - \omega L^H) \\ &= (D + \omega L)^{-1}(D + \omega L - \omega A) \\ &= I - (\omega^{-1}D + L)^{-1}A. \end{aligned}$$

If $\lambda \in \sigma(B_\omega)$, then there exists a nonzero $\mathbf{x} \in \mathbb{F}^n$ such that $B_\omega\mathbf{x} = \lambda\mathbf{x}$, and some straightforward algebraic manipulations give

$$\frac{1}{1 - \lambda} = \frac{\mathbf{x}^H(\omega^{-1}D + L)\mathbf{x}}{\mathbf{x}^HA\mathbf{x}}.$$

Since $\omega \in (0, 2)$ and $A > 0$, we have that $\frac{2}{\omega} > 1$, and by Exercise 4.27 we have $\mathbf{x}^HD\mathbf{x} > 0$. Moreover, since $A > 0$ we have $\mathbf{x}^HA\mathbf{x} > 0$. Therefore,

$$\Re\left(\frac{1}{1 - \lambda}\right) = \frac{\mathbf{x}^H(2\omega^{-1}D + L + L^H)\mathbf{x}}{2\mathbf{x}^HA\mathbf{x}} = \frac{1}{2} \left(\left(\frac{2}{\omega} - 1\right) \frac{\mathbf{x}^HD\mathbf{x}}{\mathbf{x}^HA\mathbf{x}} + 1 \right) > \frac{1}{2}.$$

Hence, for $\lambda = u + iv$, we have that

$$\frac{1}{2} < \Re\left(\frac{1}{1 - \lambda}\right) = \Re\left(\frac{1}{1 - \lambda} \cdot \frac{1 - \bar{\lambda}}{1 - \bar{\lambda}}\right) = \Re\left(\frac{1 - \bar{\lambda}}{1 - 2\Re(\lambda) + |\lambda|^2}\right) = \frac{1 - u}{1 - 2u + u^2 + v^2}.$$

Cross multiplying gives $|\lambda|^2 = u^2 + v^2 < 1$. $\square$

13.2 Minimal Polynomials and Krylov Subspaces

In this section and the next, we consider *Krylov methods*, which are related to Newton's method, and which form the basis of many iterative methods for sparse linear algebra problems. The basic idea is to construct a sequence of low-dimensional subspaces in which an approximation to the true solution can be solved somewhat easily. If the subspaces are well chosen, the true solution can be closely approximated in relatively few steps.

Before diving into Krylov spaces, we define *minimal polynomials*, which are similar to characteristic polynomials, and which provide the theoretical foundation for Krylov methods. We then show how Krylov subspaces can be used to approximate the solutions of linear systems. In the next section, we show how the algorithms work in practice.

13.2.1 Minimal Polynomials

The Cayley–Hamilton theorem (see Corollary 12.4.8) states that the characteristic polynomial $p(z)$ of a given matrix $A \in M_n(\mathbb{F})$ annihilates the matrix, that is, $p(A) = 0$. But it often happens that a polynomial of smaller degree will also annihilate A .

Definition 13.2.1. The minimal polynomial $\hat{p} \in \mathbb{F}[x]$ of a matrix $A \in M_n(\mathbb{F})$ is the monic, nonzero polynomial of least degree in $\mathbb{F}[z]$ such that $\hat{p}(A) = 0$.

Proposition 13.2.2. Given $A \in M_n(\mathbb{F})$, there exists a unique minimal polynomial.

Proof. See Exercise 13.6. $\square$

Lemma 13.2.3. Let $\hat{p} \in \mathbb{F}[x]$ be the minimal polynomial of a matrix $A \in M_n(\mathbb{F})$. If λ is an eigenvalue of A , then λ is a root of $\hat{p}$; that is, $\hat{p}(\lambda) = 0$.

Proof. By the spectral mapping theorem (Theorem 12.7.1), if λ is an eigenvalue of A , then $\hat{p}(\lambda)$ is an eigenvalue of $\hat{p}(A)$. Hence for some $\mathbf{x} \neq \mathbf{0}$, we have that $\hat{p}(A)\mathbf{x} = \hat{p}(\lambda)\mathbf{x}$, but since $\hat{p}(A) = 0$, we must also have that $\hat{p}(\lambda) = 0$. $\square$

Lemma 13.2.4. Let $\hat{p} \in \mathbb{F}[x]$ be the minimal polynomial of $A \in M_n(\mathbb{F})$. If $q \in \mathbb{F}[x]$ satisfies $q(A) = 0$, then $\hat{p}(z)$ divides $q(z)$.

Proof. Since $q(A) = 0$, the degree of q is greater than or equal to the degree of $\hat{p}$. Using polynomial division,⁴⁹ we can write $q = m\hat{p} + r$, where $m, r \in \mathbb{F}[x]$ and $\deg(r) < \deg(\hat{p})$. Thus, $r(A) = q(A) - m(A)\hat{p}(A) = 0$, which contradicts the minimality of $\hat{p}$, unless $r = 0$. $\square$

The minimal polynomial can be constructed from the spectral decomposition, as described in the next proposition.

⁴⁹We assume the reader has seen polynomial division before. For a review and for more about polynomial division see Definition 15.2.1 and Theorem 15.2.4.

Proposition 13.2.5. *Given $A \in M_n(\mathbb{F})$, let $A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda$ be the spectral decomposition. For each distinct eigenvalue λ , let m_λ denote the order of nilpotency of D_λ . In this case, the minimal polynomial of A is*

$$\hat{p}(z) = \prod_{\lambda \in \sigma(A)} (z - \lambda)^{m_\lambda}. \quad (13.8)$$

Proof. The proof is Exercise 13.7. $\square$

Example 13.2.6. In practice the characteristic polynomial and the minimal polynomial are often the same, but they can differ. Consider the matrix

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

It is easy to see that the characteristic polynomial is $p(z) = (z - 3)^4$, whereas the minimal polynomial is $\hat{p}(z) = (z - 3)^2$.

13.2.2 Krylov Solution for Nonsingular Matrices

Let $\hat{p}(z) = \sum_{k=0}^d c_k z^k$ be the minimal polynomial of a matrix $A \in M_n(\mathbb{F})$. If A is nonsingular, then c_0 is nonzero, and we can write (see Exercise 13.9)

$$A^{-1} = -\frac{1}{c_0} \sum_{k=0}^{d-1} c_{k+1} A^k. \quad (13.9)$$

Applying this to the linear system $A\mathbf{x} = \mathbf{b}$, we have

$$\mathbf{x} = A^{-1}\mathbf{b} = -\frac{1}{c_0} \sum_{k=0}^{d-1} c_{k+1} A^k \mathbf{b}. \quad (13.10)$$

This implies that the solution $\mathbf{x}$ lies in the d -dimensional subspace spanned by the vectors $\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{d-1}\mathbf{b}$.

Definition 13.2.7. *For $A \in M_n(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$, the k th Krylov subspace of A generated by $\mathbf{b}$ is*

$$\mathcal{K}_k(A, \mathbf{b}) = \text{span}\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{k-1}\mathbf{b}\}.$$

If $\dim(\mathcal{K}_k(A, \mathbf{b})) = k$, we call $\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{k-1}\mathbf{b}\}$ the Krylov basis of $\mathcal{K}_k(A, \mathbf{b})$.

Remark 13.2.8. From (13.10) it follows that the solution of the invertible linear system $A\mathbf{x} = \mathbf{b}$ lies in the Krylov subspace $\mathcal{K}_d(A, \mathbf{b})$.

Remark 13.2.9. Observe that

$$\mathcal{K}_1(A, \mathbf{b}) \subset \mathcal{K}_2(A, \mathbf{b}) \subset \mathcal{K}_3(A, \mathbf{b}) \subset \cdots$$

The fundamental idea of Krylov subspace methods is to search for approximate solutions of the linear system $A\mathbf{x} = \mathbf{b}$ in each of the low-dimensional subspaces $\mathcal{K}_k(A, \mathbf{b})$, where $k = 1, 2, \dots$, until the approximate solution is close enough to the true solution that we can terminate the algorithm. While we can prove that $\mathcal{K}_d(A, \mathbf{b})$ contains the exact solution, we can usually get a sufficiently close approximation with a much smaller value of k .

Remark 13.2.10. It's clear that $\mathcal{K}_k(A, \mathbf{b}) \subset \mathcal{K}_d(A, \mathbf{b})$ whenever $k \leq d$, but this also holds for $k > d$, in particular, given a polynomial $q \in \mathbb{F}[x]$ of any order, we can write it in the form $q = m\hat{p} + r$, where $m, r \in \mathbb{F}[x]$ and $\deg(r) < \deg(p)$. Thus, $q(A)\mathbf{b} = m(A)p(A)\mathbf{b} + r(A)\mathbf{b} = r(A)\mathbf{b} \in \mathcal{K}_d(A, \mathbf{b})$.

Example 13.2.11. Suppose

$$A = \begin{bmatrix} 2 & 3 & 5 & 7 \\ 0 & 2 & 3 & 5 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The minimal polynomial and is given by

$$\hat{p}(z) = (z - 2)^3(z - 3) = z^4 - 9z^3 + 30z^2 - 44z + 24$$

and the 4th Krylov subspace of A generated by $\mathbf{b}$ is given by

$$\mathcal{K}_4(A, \mathbf{b}) = \text{span} \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 17 \\ 10 \\ 5 \\ 3 \end{bmatrix}, \begin{bmatrix} 110 \\ 50 \\ 19 \\ 9 \end{bmatrix}, \begin{bmatrix} 528 \\ 202 \\ 65 \\ 27 \end{bmatrix} \right\}.$$

We may write the solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$ in the Krylov basis as

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{-1}{24} (A^3\mathbf{b} - 9A^2\mathbf{b} + 30A\mathbf{b} - 44\mathbf{b}) = \frac{1}{6} \begin{bmatrix} -1 \\ -2 \\ 0 \\ 2 \end{bmatrix}.$$

13.2.3 Krylov Solution for Singular Matrices

Definition 13.2.12. Let $A \in M_n(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$. The linear system $A\mathbf{x} = \mathbf{b}$ is said to have a Krylov solution if for some positive integer k there exists $\mathbf{x} \in \mathcal{K}_k(A, \mathbf{b})$ such that $A\mathbf{x} = \mathbf{b}$.

In the previous subsection we saw that $A\mathbf{x} = \mathbf{b}$ always has a Krylov solution if A is nonsingular. We now consider the case that $A \in M_n(\mathbb{F})$ is singular.

Remark 13.2.13. Let $A \in M_n(\mathbb{F})$, and let d be the degree of the minimal polynomial of A . By Remark 13.2.10, the linear system $A\mathbf{x} = \mathbf{b}$ has a Krylov solution if and only if there exists $\mathbf{x} \in \mathcal{K}_d(A, \mathbf{b})$ satisfying $A\mathbf{x} = \mathbf{b}$.

Lemma 13.2.14. If $N \in M_n(\mathbb{F})$ is nilpotent of order d and $\mathbf{b} \neq \mathbf{0}$, then $N\mathbf{x} = \mathbf{b}$ has no Krylov solution. In other words, for any $\mathbf{x} \in \mathcal{K}_k(N, \mathbf{b})$ we cannot have $N\mathbf{x} = \mathbf{b}$.

Proof. Since $N^{d-1} \neq 0$ and $N^d = 0$, the minimal polynomial of N has degree d . If $\mathbf{x} \in \mathcal{K}_d(N, \mathbf{b})$ satisfies $N\mathbf{x} = \mathbf{b}$, then $\mathbf{x}$ may be written as a linear combination of the Krylov basis, as

$$\mathbf{x} = \sum_{j=0}^{m-1} a_j N^j \mathbf{b} = a_0 \mathbf{b} + a_1 N \mathbf{b} + \cdots + a_{m-2} N^{m-2} \mathbf{b} + a_{m-1} N^{m-1} \mathbf{b}.$$

But this implies

$$\mathbf{b} = N\mathbf{x} = a_0 N \mathbf{b} + a_1 N^2 \mathbf{b} + \cdots + a_{m-2} N^{m-1} \mathbf{b},$$

and hence

$$(I - a_0 N - a_1 N^2 - \cdots - a_{m-2} N^{m-1}) \mathbf{b} = \mathbf{0}. \quad (13.11)$$

The matrix in parentheses is nonsingular because it can be written in the form $I + M$, where $M = -a_0 N - a_1 N^2 - \cdots - a_{m-2} N^{m-1}$ is nilpotent; see Exercise 13.11(i). Hence $\mathbf{b} = \mathbf{0}$, which is a contradiction $\square$

Theorem 13.2.15 (Existence of a Krylov Solution). Given $A \in M_n(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$, the linear system $A\mathbf{x} = \mathbf{b}$ has a Krylov solution if and only if $\mathbf{b} \in \mathcal{R}(A^m)$, where $m = \text{ind}(A)$. Moreover, if a Krylov solution exists, then it is unique and can be written as

$$\mathbf{x} = A^D \mathbf{b}. \quad (13.12)$$

Proof. The nonsingular case is immediate. Assume now that A is singular. Let $A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda + D_\lambda$ be the spectral decomposition of A and let $P_* = \sum_{\lambda \neq 0} P_\lambda$.

If $\mathbf{x}$ is a Krylov solution, then by Remark 13.2.13, it follows that $\mathbf{x} \in \mathcal{K}_d(A, \mathbf{b})$, where d is the degree of the minimal polynomial of A . Thus, we can write $\mathbf{x}$ as a linear combination $\mathbf{x} = \sum_{j=0}^{d-1} a_j A^j \mathbf{b}$. Left multiplying by P_* and P_0 gives

$$P_* \mathbf{x} = \sum_{j=0}^{d-1} a_j A^j P_* \mathbf{b} \quad \text{and} \quad P_0 \mathbf{x} = \sum_{j=0}^{d-1} a_j D_0^j P_0 \mathbf{b}, \quad (13.13)$$

and hence $P_* \mathbf{x} \in \mathcal{K}_d(A, P_* \mathbf{b})$ and $P_0 \mathbf{x} \in \mathcal{K}_d(D_0, P_0 \mathbf{b})$.

Using the spectral decomposition of A , it is straightforward to verify that $P_0 A = D_0$, and thus since $A\mathbf{x} = \mathbf{b}$, then $P_0 \mathbf{x}$ is a solution of the nilpotent linear system $D_0 P_0 \mathbf{x} = P_0 \mathbf{b}$, which implies $P_0 \mathbf{b} = \mathbf{0}$ by the lemma. Therefore, $\mathbf{b} = P_* \mathbf{b}$, which implies that $\mathbf{b} \in \mathcal{R}(P_*) = \mathcal{R}(A^m)$ and $\mathbf{x} = A^D \mathbf{b}$ by Exercise 12.48.

Conversely, if $\mathbf{b} \in \mathcal{R}(A^m)$, then by Exercise 12.48 we have $AA^D\mathbf{b} = \mathbf{b}$, which implies that $\mathbf{x} = A^D\mathbf{b}$ is a solution to $A\mathbf{x} = \mathbf{b}$. To prove that $A^D\mathbf{b} \in \mathcal{K}_k(A, \mathbf{b})$, for some $k \in \mathbb{N}$, we observe that A is invertible when restricted to the subspace $\mathcal{R}(P_*)$, and its inverse is A^D (this follows since $AA^D = A^DA = P_*$). Thus A^DP_* can be written as a polynomial q in AP_* via (13.9), and so it follows that

$$\mathbf{x} = A^D\mathbf{b} = A^DP_*\mathbf{b} = q(AP_*)\mathbf{b} = q(A)P_*\mathbf{b} = q(A)\mathbf{b}.$$

To prove uniqueness, suppose there exists another solution $\mathbf{y} = \sum_{j=0}^k c_j A^j \mathbf{b}$ of the linear system. It follows that $\mathbf{y} = \mathbf{x} + \mathbf{z}$ for some $\mathbf{z} \in \mathcal{N}(A)$. Note that $P_*\mathbf{y} = P_*\mathbf{x}$ and

$$P_0\mathbf{y} = \sum_{j=0}^k c_j P_0 A^j \mathbf{b} = \sum_{j=0}^k c_j D_0^j P_0 \mathbf{b} = \sum_{j=0}^k c_j D_0^j \mathbf{0} = \mathbf{0}.$$

Thus, $\mathbf{y} = P_*\mathbf{y} + P_0\mathbf{y} = P_*\mathbf{y} = P_*\mathbf{x} = A^D\mathbf{b}$. $\square$

Remark 13.2.16. By Remark 12.9.4, we can write the Drazin inverse of A as

$$A^D = S^{-1} \begin{bmatrix} C_*^{-1} & 0 \\ 0 & 0 \end{bmatrix} S.$$

Since C_* is invertible, we can write C_*^{-1} as a linear combination of powers of C_* , that is, as a polynomial in C_* (see 13.10). The degree of this polynomial, which is also the dimension of the largest Krylov subspace $\mathcal{K}_k(A, \mathbf{b})$, is degree d of the minimal polynomial of A minus the index m of A , that is, $k = d - m$.

13.3 The Arnoldi Iteration and GMRES Methods

Although the solution $\mathbf{x}$ to a linear system $A\mathbf{x} = \mathbf{b}$ lies in the span of the Krylov basis $\{\mathbf{b}, A\mathbf{b}, \dots, A^{d-1}\mathbf{b}, A^{d-1}\mathbf{b}\}$, the Krylov basis is not suitable for numerical computations. This is because as k becomes large, the vectors $A^{k-1}\mathbf{b}$ and $A^k\mathbf{b}$ become nearly parallel, which causes the basis to become ill conditioned; that is, small errors and round-off effects tend to compound into large errors in the solution.

Arnoldi iteration is a variant of the (modified) Gram-Schmidt orthonormalization process (see Theorem 3.3.1) that constructs an orthonormal set of vectors spanning the Krylov subspace $\mathcal{K}_d(A, \mathbf{b})$. This new orthonormal basis is chosen so that the matrix representation of A in terms of this basis is nearly upper triangular (called upper Hessenberg), which also makes solving many linear algebra problems relatively easy.

In this section, we use Arnoldi iteration to solve linear systems using what is called the *generalized minimal residual method*, or GMRES⁵⁰ for short. For each $k \in \mathbb{Z}^+$, the GMRES algorithm determines the best approximate solution $\mathbf{x}_k \in \mathcal{K}_k(A, \mathbf{b})$ of the linear system $A\mathbf{x} = \mathbf{b}$ and arrives at the exact solution once $k = d$. In practice, however, we can usually get a good approximation with far fewer iterations.

⁵⁰Pronounced G-M-Rez.

There are several linear solvers that use Arnoldi iteration. GMRES is fairly general and serves as a good example. For many systems, it will perform better than Gaussian elimination (see Section 2.7), or, more precisely, the LU decomposition (see Application 2.7.15). It is also often better than the methods described in Section 13.1. Generally speaking, GMRES performs well when the matrix A is sparse and when the eigenvalues of A are not clustered around the origin.

13.3.1 Hessenberg Matrices and the Arnoldi Iteration

Definition 13.3.1. A Hessenberg matrix is a matrix $A \in M_{m \times n}(\mathbb{F})$ with all zeros below the first subdiagonal; that is, if $A = [a_{i,j}]$, then $a_{i,j} = 0$ whenever $i > j + 1$.

Example 13.3.2. Every 2×2 matrix is automatically Hessenberg. Hessenberg 3×3 matrices must have a zero in the bottom left corner, and Hessenberg 4×3 matrices have three zeros in the bottom left corner. We denote these as

$$\begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \end{bmatrix} \in M_3(\mathbb{F}) \quad \text{and} \quad \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \in M_{4 \times 3}(\mathbb{F}),$$

where $*$ indicates an arbitrary entry—not necessarily zero or nonzero.

The Arnoldi process begins with an initial vector $\mathbf{b} \neq \mathbf{0}$. Since the basis being constructed is orthonormal, the first step is to normalize $\mathbf{b}$ to get $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|$. This is the Arnoldi basis for the initial Krylov subspace $\mathcal{K}_1(A, \mathbf{b})$. Proceeding inductively, if the orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_j$ for $\mathcal{K}_j(A, \mathbf{b})$ has been computed, then construct the next basis element by subtracting from $A\mathbf{q}_j \in \mathcal{K}_{j+1}(A, \mathbf{b})$ its projection onto $\mathcal{K}_j(A, \mathbf{b})$. This goes as follows. First, set

$$h_{i,j} = \langle \mathbf{q}_i, A\mathbf{q}_j \rangle = \mathbf{q}_i^H A\mathbf{q}_j, \quad i \leq j.$$

Now set

$$\hat{\mathbf{q}}_{j+1} = A\mathbf{q}_j - \sum_{i=1}^j h_{i,j} \mathbf{q}_i \quad \text{and} \quad h_{j+1,j} = \|\hat{\mathbf{q}}_{j+1}\|. \quad (13.14)$$

If $h_{j+1,j} = 0$, then $\|\hat{\mathbf{q}}_{j+1}\| = 0$, which implies $A\mathbf{q}_j \in \mathcal{K}_j(A, \mathbf{b})$, and the algorithm terminates. Otherwise set

$$\mathbf{q}_{j+1} = \frac{\hat{\mathbf{q}}_{j+1}}{h_{j+1,j}}. \quad (13.15)$$

This algorithm produces an orthonormal basis for each $\mathcal{K}_k(A, \mathbf{b})$ for $k = 1, 2, \dots, m$, where m is the smallest integer such that $\mathcal{K}_m(A, \mathbf{b}) = \mathcal{K}_{m+1}(A, \mathbf{b})$.

For the ambient space $\mathbb{F}^n$, define for each $k = 1, 2, \dots, m$ the $n \times k$ matrix Q_k whose j th column is the j th Arnoldi basis vector $\mathbf{q}_j$. Let $\tilde{H}_k$ be the $(k+1) \times k$ Hessenberg matrix whose (i, j) entry is h_{ij} . Note that when $i > j + 1$,

we haven't yet defined $h_{i,j}$, so we just set it to 0; this is why the matrix $\tilde{H}_k$ is Hessenberg.

Example 13.3.3. The first two Hessenberg matrices $\tilde{H}_1$ and $\tilde{H}_2$ are

$$\tilde{H}_1 = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^H A \mathbf{q}_1 \\ \|\hat{\mathbf{q}}_2\| \end{bmatrix}$$

and

$$\tilde{H}_2 = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ 0 & h_{23} \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1^H A \mathbf{q}_1 & \mathbf{q}_1^H A \mathbf{q}_2 \\ \|\hat{\mathbf{q}}_2\| & \mathbf{q}_2^H A \mathbf{q}_2 \\ 0 & \|\hat{\mathbf{q}}_3\| \end{bmatrix}.$$

We can combine (13.14) and (13.15) as

$$A\mathbf{q}_j = h_{1,j}\mathbf{q}_1 + \cdots + h_{j,j}\mathbf{q}_j + h_{j+1,j}\mathbf{q}_{j+1}$$

or, in matrix form,

$$AQ_j = Q_{j+1}\tilde{H}_j. \quad (13.16)$$

Note that since the columns of Q_j are orthonormal and equal to the first j columns of Q_{j+1} , the product $Q_j^H Q_{j+1}$ is the $j \times j$ identity matrix with an extra column of zeros adjoined to the right. That is, it is the orthogonal projection that forgets the last coordinate. So multiplying (13.16) on the left by Q_j^H gives

$$Q_j^H A Q_j = H_j, \quad (13.17)$$

where H_j is the matrix obtained from $\tilde{H}_j$ by removing the bottom row. Note that in the special case that $j = n$, the matrix Q_n is square and orthonormal.

Example 13.3.4. Consider the linear system $A\mathbf{x} = \mathbf{b}$ from Example 13.2.11, that is,

$$A = \begin{bmatrix} 2 & 3 & 5 & 7 \\ 0 & 2 & 3 & 5 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Applying the Arnoldi method yields the following sequence of orthonormal and Hessenberg matrices computed to four digits of accuracy:

$$Q_1 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad \tilde{H}_1 = \begin{bmatrix} 8.75 \\ 5.403 \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} 0.5 & 0.7635 \\ 0.5 & 0.1157 \\ 0.5 & -0.3471 \\ 0.5 & -0.5322 \end{bmatrix}, \quad \tilde{H}_2 = \begin{bmatrix} 8.75 & -5.472 \\ 5.403 & -1.495 \\ 0 & 0.7238 \end{bmatrix},$$

$$Q_3 = \begin{bmatrix} 0.5 & 0.7635 & 0.4025 \\ 0.5 & 0.1157 & -0.7759 \\ 0.5 & -0.3471 & -0.1016 \\ 0.5 & -0.5322 & 0.4750 \end{bmatrix}, \quad \tilde{H}_3 = \begin{bmatrix} 8.75 & -5.472 & 2.230 \\ 5.403 & -1.495 & -0.1341 \\ 0 & 0.7238 & 0.6715 \\ 0 & 0 & 0.1634 \end{bmatrix},$$

and

$$Q_4 = \begin{bmatrix} 0.5 & 0.7635 & 0.4025 & 0.0710 \\ 0.5 & 0.1157 & -0.7759 & -0.3668 \\ 0.5 & -0.3471 & -0.1016 & 0.7869 \\ 0.5 & -0.5322 & 0.4750 & -0.4911 \end{bmatrix},$$

$$H_4 = \begin{bmatrix} 8.75 & -5.472 & 2.230 & -1.331 \\ 5.403 & -1.495 & -0.1341 & 0.3009 \\ 0 & 0.7238 & 0.6715 & -0.2531 \\ 0 & 0 & 0.1634 & 1.074 \end{bmatrix}$$

13.3.2 The GMRES Algorithm

GMRES is based on the observation that the solution $\mathbf{x}$ of the linear system $A\mathbf{x} = \mathbf{b}$ can be found by solving the least squares problem, that is, minimizing the residual $\|A\mathbf{x} - \mathbf{b}\|_2$. To find this minimizer, the algorithm constructs successive approximations $\mathbf{x}_k$ of $\mathbf{x}$ by searching for the vector $\mathbf{y} \in \mathcal{K}_k(A, \mathbf{b})$ such that $\|A\mathbf{y} - \mathbf{b}\|_2$ is minimized. In other words, for each $k \in \mathbb{N}$, let

$$\mathbf{x}_k = \operatorname{argmin}_{\mathbf{y} \in \mathcal{K}_k(A, \mathbf{b})} \|A\mathbf{y} - \mathbf{b}\|_2.$$

Once $\mathcal{K}_k(A, \mathbf{b}) = \mathcal{K}_{k+1}(A, \mathbf{b})$, we have that $\mathbf{x} \in \mathcal{K}_k(A, \mathbf{b})$, and so the least squares solution is achieved; that is, $\mathbf{x}_k = \mathbf{x}$.

Using the Arnoldi basis $\mathbf{q}_1, \dots, \mathbf{q}_k$ of $\mathcal{K}_k(A, \mathbf{b})$ described above, any $\mathbf{y} \in \mathcal{K}_k(A, \mathbf{b})$ can be written as $\mathbf{y} = Q_k \mathbf{z}$ for some $\mathbf{z} \in \mathbb{F}^k$, where Q_k is the matrix of column vectors $\mathbf{q}_1, \dots, \mathbf{q}_k$. It follows that

$$\|A\mathbf{y} - \mathbf{b}\|_2 = \|AQ_k \mathbf{z} - \mathbf{b}\|_2 = \|Q_{k+1} \tilde{H}_k \mathbf{z} - \mathbf{b}\|_2,$$

where $\tilde{H}_k$ is the $(k+1) \times k$ upper-Hessenberg matrix described in (13.16). Using Exercise 4.31, we have

$$\|Q_{k+1} \tilde{H}_k \mathbf{z} - \mathbf{b}\|_2 = \|\tilde{H}_k \mathbf{z} - Q_{k+1}^H \mathbf{b}\|_2 = \|\tilde{H}_k \mathbf{z} - \beta \mathbf{e}_1\|_2,$$

where $\beta = \|\mathbf{b}\|_2$. Thus, in the k th iteration of GMRES the problem is to find $\mathbf{z}_k \in \mathbb{F}^k$ minimizing $\|\tilde{H}_k \mathbf{z} - \beta \mathbf{e}_1\|_2$. We write this as

$$\mathbf{z}_k = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}^k} \|\tilde{H}_k \mathbf{z} - \beta \mathbf{e}_1\|_2 \quad (13.18)$$

and define the k th approximate solution of the linear system as $\mathbf{x}_k = Q_k \mathbf{z}_k$.

Once the Arnoldi basis for step k is known, the solution of the minimization problem is relatively fast, both because the special structure of H_k allows us to use previous solutions to help find the next solution and because k is usually much smaller than n .

To complete the solution, we use the QR decomposition of $\tilde{H}_k$. Recall the QR decomposition (Section 3.3.3), where $\tilde{H}_k \in M_{(k+1) \times k}(\mathbb{F})$ can be written as a product

$$\tilde{H}_k = \tilde{\Omega}_k \tilde{R}_k, \quad (13.19)$$

where $\tilde{\Omega}_k \in M_{k+1}(\mathbb{F})$ is orthonormal (that is, satisfies $\tilde{\Omega}_k^H \tilde{\Omega}_k = \tilde{\Omega}_k \tilde{\Omega}_k^H = I_{k+1}$) and $\tilde{R}_k \in M_{(k+1) \times k}(\mathbb{F})$ is upper triangular. Thus,

$$\mathbf{z}_k = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}^k} \left\| \tilde{H}_k \mathbf{z} - \beta \mathbf{e}_1 \right\|_2 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}^k} \left\| \tilde{R}_k \mathbf{z} - \beta \tilde{\Omega}_k^H \mathbf{e}_1 \right\|_2. \quad (13.20)$$

Since $\tilde{R}_k$ is upper triangular, the linear system $\tilde{R}_k \mathbf{z} = \beta \tilde{\Omega}_k^H \mathbf{e}_1$ can be solved quickly by back substitution. In fact, since $\tilde{R}_k$ is a $(k+1) \times k$ upper-triangular matrix, its bottom row is all zeros. Hence, we write

$$\tilde{R}_k = \begin{bmatrix} R_k \\ 0 \end{bmatrix},$$

where R_k is a $k \times k$ upper-triangular matrix. We can also split off the last entry of the right-hand side $\beta \tilde{\Omega}_k^H \mathbf{e}_1$. Setting

$$\beta \tilde{\Omega}_k^H \mathbf{e}_1 = \begin{bmatrix} \mathbf{g}_k \\ \gamma_k \end{bmatrix} \quad (13.21)$$

gives the solution of (13.18) as that of the linear system $R_k \mathbf{z}_k = \mathbf{g}_k$, which can be computed via back substitution.

Remark 13.3.5. The QR decomposition, followed by back substitution, is not faster than row reduction for solving linear systems, and so at first blush, it might seem that this approach of solving a linear system at each step is ridiculous. However, as we see below, the special structure of the upper-Hessenberg matrix $\tilde{H}_k$ makes it easy to solve these systems recursively, and thus this method solves the overall problem rather efficiently.

Now compute the least squares solution for the next iterate. Note that

$$\tilde{H}_{k+1} = \begin{bmatrix} \tilde{H}_k & \mathbf{h}_{k+1} \\ 0 & h_{k+2,k+1} \end{bmatrix}, \quad (13.22)$$

where $\mathbf{h}_{k+1} = [h_{1,k+1} \ h_{2,k+1} \ \dots \ h_{k+1,k+1}]^T$. Left multiplying (13.22) by the Hermitian matrix

$$\begin{bmatrix} \tilde{\Omega}_k^H & 0 \\ 0 & 1 \end{bmatrix}$$

yields

$$\begin{bmatrix} \tilde{\Omega}_k^H & 0 \\ 0 & 1 \end{bmatrix} \tilde{H}_{k+1} = \begin{bmatrix} \tilde{R}_k & \Omega_k^H \mathbf{h}_{k+1} \\ 0 & h_{k+2,k+1} \end{bmatrix} = \begin{bmatrix} R_k & \mathbf{r}_{k+1} \\ 0 & \rho \\ 0 & \sigma \end{bmatrix}. \quad (13.23)$$

If $\sigma = 0$, then the rightmost matrix of the above expression is $\tilde{R}_{k+1}$, which is upper triangular, and the solution of the linear system $\tilde{R}_{k+1} \mathbf{z} = \beta \tilde{\Omega}_{k+1}^H \mathbf{e}_1$ can be found quickly with back substitution.

If, instead, we have $\sigma \neq 0$, then perform a special rotation to eliminate the bottom row. Specifically, let

$$G_k = \begin{bmatrix} I_k & 0 & 0 \\ 0 & c_k & s_k \\ 0 & -s_k & c_k \end{bmatrix}, \quad (13.24)$$

where

$$c_k = \frac{\rho}{\sqrt{\rho^2 + \sigma^2}} \quad \text{and} \quad s_k = \frac{\sigma}{\sqrt{\rho^2 + \sigma^2}}.$$

Thus,

$$\underbrace{G_k \begin{bmatrix} \tilde{\Omega}_k^H & 0 \\ 0 & 1 \end{bmatrix}}_{\tilde{\Omega}_{k+1}^H} \tilde{H}_{k+1} = \begin{bmatrix} I_k & 0 & 0 \\ 0 & c_k & s_k \\ 0 & -s_k & c_k \end{bmatrix} \begin{bmatrix} R_k & \mathbf{r}_{k+1} \\ 0 & \rho \\ 0 & \sigma \end{bmatrix} = \begin{bmatrix} R_k & \mathbf{r}_{k+1} \\ 0 & r_{k+1,k+1} \\ 0 & 0 \end{bmatrix} = \tilde{R}_{k+1},$$

where $r_{k+1,k+1} = \sqrt{\rho^2 + \sigma^2}$. This gives the next iterate of (13.19) without having to compute the QR decomposition again. In particular, we have $\tilde{H}_{k+1} = \tilde{\Omega}_{k+1} \tilde{R}_{k+1}$, where $\tilde{\Omega}_{k+1}$ is orthonormal (since it is the product of two orthonormal matrices) and $\tilde{R}_{k+1}$ is upper triangular.

We can also split up the next iterate of (13.21). Note that

$$\beta \tilde{\Omega}_{k+1}^H \mathbf{e}_1 = \beta G_k \begin{bmatrix} \tilde{\Omega}_k^H & 0 \\ 0 & 1 \end{bmatrix} \mathbf{e}_1 = G_k \begin{bmatrix} \beta \tilde{\Omega}_k^H \mathbf{e}_1 \\ 0 \end{bmatrix} = G_k \begin{bmatrix} \mathbf{g}_k \\ \gamma_k \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_k \\ c_k \gamma_k \\ -s_k \gamma_k \end{bmatrix}.$$

It follows that

$$\|\tilde{H}_{k+1} \mathbf{z} - \beta \mathbf{e}_1\|_2 = \|\tilde{R}_{k+1} \mathbf{z} - \beta \tilde{\Omega}_{k+1}^H \mathbf{e}_1\|_2 = \left\| \begin{bmatrix} R_k & \mathbf{r}_{k+1} \\ 0 & r_{k+1,k+1} \\ 0 & 0 \end{bmatrix} \mathbf{z} - \begin{bmatrix} \mathbf{g}_k \\ c_k \gamma_k \\ -s_k \gamma_k \end{bmatrix} \right\|_2.$$

Hence, the solution $\mathbf{z}_{k+1}$ is found by solving the linear system

$$\begin{bmatrix} R_k & \mathbf{r}_{k+1} \\ 0 & r_{k+1,k+1} \end{bmatrix} \mathbf{z}_{k+1} = \begin{bmatrix} \mathbf{g}_k \\ c_k \gamma_k \end{bmatrix}, \quad (13.25)$$

which can be found by back substitution.

Remark 13.3.6. Theorem 13.2.15, combined with Remark 13.2.16, guarantees that GMRES will converge in no more than d iterations, where d is the degree of the minimal polynomial minus the index of A . But, as mentioned earlier, GMRES can often get a sufficiently good approximation to the solution with far fewer iterates. Since the least squares error is given by $|\gamma_k|$, we can monitor the error at each iteration and terminate when it is sufficiently small.

13.3.3 A Worked Example

We conclude this section with a worked example of the GMRES method. Consider the linear system from Examples 13.2.11 and 13.3.4. Although this system is easy

to solve with back substitution (since it's upper triangular), we solve it here with GMRES for illustrative reasons. In what follows we provide four digits of accuracy.

Since each $\tilde{H}_k$ is given in Example 13.3.4, we can solve each iteration fairly easily. For $k = 1$, it follows from (13.20) that

$$\mathbf{z}_1 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}} \left\| \tilde{H}_1 \mathbf{z} - \beta \mathbf{e}_1 \right\|_2 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}} \left\| \tilde{R}_1 \mathbf{z} - \beta \tilde{\Omega}_1^H \mathbf{e}_1 \right\|_2.$$

The QR decomposition of $\tilde{H}_1 = \tilde{\Omega}_1 \tilde{R}_1$ is given by

$$\tilde{H}_1 = \begin{bmatrix} 8.75 \\ 5.403 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.8509 & -0.5254 \\ 0.5254 & 0.8509 \end{bmatrix}}_{\tilde{\Omega}_1} \underbrace{\begin{bmatrix} 10.28 \\ 0 \end{bmatrix}}_{\tilde{R}_1}.$$

Thus,

$$\mathbf{z}_1 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}} \left\| \tilde{R}_1 \mathbf{z} - \beta \tilde{\Omega}_1^H \mathbf{e}_1 \right\|_2 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}} \left\| \begin{bmatrix} 10.28 \\ 0 \end{bmatrix} \mathbf{z} - \begin{bmatrix} 1.702 \\ -1.051 \end{bmatrix} \right\|_2 = 0.1655.$$

The least squares error is 1.051. For the next iteration, we have that

$$\mathbf{z}_2 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}^2} \left\| \tilde{H}_2 \mathbf{z} - \beta \mathbf{e}_1 \right\|_2 = \operatorname{argmin}_{\mathbf{z} \in \mathbb{F}^2} \left\| \tilde{R}_2 \mathbf{z} - \beta \tilde{\Omega}_2^H \mathbf{e}_1 \right\|_2.$$

Here (13.23) gives $\tilde{H}_2 = \tilde{\Omega}_2 \tilde{R}_2$,

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9113 & 0.4116 \\ 0 & -0.4116 & 0.9113 \end{bmatrix},$$

$$\begin{aligned} \tilde{\Omega}_2 = \begin{bmatrix} \tilde{\Omega}_1 & 0 \\ 0 & 1 \end{bmatrix} G_2^H &= \begin{bmatrix} 0.8509 & -0.5254 & 0 \\ 0.5254 & 0.8509 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9113 & -0.4116 \\ 0 & 0.4116 & 0.9113 \end{bmatrix} \\ &= \begin{bmatrix} 0.8509 & -0.4788 & 0.2163 \\ 0.5254 & 0.7754 & -0.3502 \\ 0 & 0.4116 & 0.9113 \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} \tilde{R}_2 = G_2 \begin{bmatrix} R_1 & \mathbf{r}_2 \\ 0 & r_{2,2} \\ 0 & 0 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9113 & 0.4116 \\ 0 & -0.4116 & 0.9113 \end{bmatrix} \begin{bmatrix} 10.28 & -5.441 \\ 0 & 1.603 \\ 0 & 0.7238 \end{bmatrix} \\ &= \begin{bmatrix} 10.28 & -5.441 \\ 0 & 1.758 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Moreover, we have that

$$\beta \tilde{\Omega}_2^H \mathbf{e}_1 = \begin{bmatrix} \mathbf{g}_1 \\ c_1 \gamma_1 \\ -s_1 \gamma_1 \end{bmatrix} = \begin{bmatrix} 1.702 \\ 0.9113(-1.051) \\ -0.4116(-1.051) \end{bmatrix} = \begin{bmatrix} 1.702 \\ -0.9576 \\ 0.4325 \end{bmatrix}.$$

Thus, the least squares solution satisfies the linear system (13.25)

$$\begin{bmatrix} 10.28 & -5.441 \\ 0 & 1.758 \end{bmatrix} \mathbf{z}_2 = \begin{bmatrix} 1.702 \\ -0.9576 \end{bmatrix},$$

which can be solved with back substitution. The solution is

$$\mathbf{z}_2 = [-0.1227 \quad -0.5446]^\top,$$

which has a least squares error of 0.4325. Repeating for $k = 3$, we have $\tilde{H}_3 = \tilde{\Omega}_3 \tilde{R}_3$, where

$$G_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.9899 & 0.1417 \\ 0 & 0 & -0.1417 & 0.9899 \end{bmatrix},$$

$$\tilde{\Omega}_3 = \begin{bmatrix} 0.8509 & -0.4788 & 0.2163 & -0.0307 \\ 0.5254 & 0.7754 & -0.3502 & 0.04965 \\ 0 & 0.4116 & 0.9021 & -0.1292 \\ 0 & 0 & 0.1417 & 0.9899 \end{bmatrix},$$

and

$$\tilde{R}_3 = \begin{bmatrix} 10.28 & -5.441 & 1.827 \\ 0 & 1.758 & -0.8953 \\ 0 & 0 & 1.153 \\ 0 & 0 & 0 \end{bmatrix}.$$

Moreover, we have that

$$\beta \tilde{\Omega}_3^H \mathbf{e}_1 = \begin{bmatrix} \mathbf{g}_2 \\ c_2 \gamma_2 \\ -s_2 \gamma_2 \end{bmatrix} = \begin{bmatrix} 1.702 \\ -0.9576 \\ 0.4281 \\ -0.06131 \end{bmatrix}.$$

Thus, the least squares solution is satisfied by the linear system (13.25)

$$\begin{bmatrix} 10.28 & -5.441 & 1.827 \\ 0 & 1.758 & -0.8953 \\ 0 & 0 & 1.153 \end{bmatrix} \mathbf{z}_2 = \begin{bmatrix} 1.702 \\ -0.9576 \\ 0.4281 \end{bmatrix}.$$

Solving with back substitution gives

$$\mathbf{z}_3 = [-0.08860 \quad -0.3555 \quad 0.3714]^\top,$$

which has a least squares error of 0.06131. Finally, repeating for $k = 4$, we have $\tilde{H}_4 = \tilde{\Omega}_4 \tilde{R}_4$, where $G_4 = I_5$,

$$\tilde{\Omega}_3 = \begin{bmatrix} 0.8509 & -0.4788 & 0.2163 & -0.0307 & 0 \\ 0.5254 & 0.7754 & -0.3502 & 0.04965 & 0 \\ 0 & 0.4116 & 0.9021 & -0.1292 & 0 \\ 0 & 0 & 0.1417 & 0.9899 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\tilde{R}_3 = \begin{bmatrix} 10.28 & -5.441 & 1.827 & -0.9746 \\ 0 & 1.758 & -0.8953 & 0.7665 \\ 0 & 0 & 1.153 & -0.4654 \\ 0 & 0 & 0 & 1.1513 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Moreover, we have that

$$\beta \tilde{\Omega}_3^H \mathbf{e}_1 = \begin{bmatrix} \mathbf{g}_2 \\ c_2 \gamma_2 \\ -s_2 \gamma_2 \end{bmatrix} = \begin{bmatrix} 1.702 \\ -0.9576 \\ 0.4281 \\ -0.06131 \\ 0 \end{bmatrix}.$$

Thus, the least squares solution is satisfied by the linear system (13.25)

$$\begin{bmatrix} 10.28 & -5.441 & 1.827 & -0.9746 \\ 0 & 1.758 & -0.8953 & 0.7665 \\ 0 & 0 & 1.153 & -0.4654 \\ 0 & 0 & 0 & 1.1513 \end{bmatrix} \mathbf{z}_2 = \begin{bmatrix} 1.702 \\ -0.9576 \\ 0.4281 \\ -0.0613 \end{bmatrix}.$$

Solving with back substitution gives

$$\mathbf{z}_3 = [-0.08860 \quad -0.3432 \quad 0.3499 \quad -0.05325]^\top,$$

which has a least squares error of 0. Since $\mathbf{x}_4 = Q_4 \mathbf{z}_4$, we have that

$$\mathbf{x}_4 = Q_4 \mathbf{z}_4 = [-0.1667 \quad -0.3333 \quad 0 \quad 0.3333]^\top \approx \frac{1}{6} [-1 \quad -2 \quad 0 \quad 2]^\top,$$

which is the solution. In other words, $\mathbf{x} = \mathbf{x}_4$.

Remark 13.3.7. In this example, we didn't stop until we got to $k = 4$ because we wanted to demonstrate that when $k = d$, we get a least squares error of zero. However, the previous step has a fairly small least squares error (0.06131). Note that

$$\mathbf{x}_3 = Q_3 \mathbf{z}_3 = [-0.1662 \quad -0.3736 \quad 0.0413 \quad 0.3213]^\top$$

is pretty close to $\mathbf{x}_4$.

13.4 * Computing Eigenvalues I

The power method (see Section 12.7.3) determines the eigenvalue of largest modulus of a given matrix $A \in M_n(\mathbb{F})$. In this section, we show how to compute the full set of eigenvalues of A , using Schur's lemma and the QR decomposition. Schur's lemma (see Section 4.4) guarantees there exists an orthonormal $Q \in M_n(\mathbb{F})$ such that $Q^H A Q$ is upper-triangular, and the diagonal elements of this upper-triangular matrix are the eigenvalues of A (see Proposition 4.1.22). We show how to compute a Schur decomposition for any matrix $A \in M_n(\mathbb{F})$ where the eigenvalues all have distinct moduli. We do this by repeatedly taking the QR decomposition in a certain clever way called *QR iteration*.

13.4.1 Uniqueness of the QR Decomposition

The QR decompositions constructed in Theorems 3.3.9 and 3.4.9 do not necessarily have only nonnegative diagonals, but they can always be rewritten that way. Specifically, if $QR = A$ is any QR decomposition of a matrix $A \in M_n(\mathbb{F})$, let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m) \in M_m(\mathbb{F})$ be the diagonal matrix whose i th entry λ_i is

$$\lambda_i = \begin{cases} \overline{\text{sign}(r_{ii})} = \bar{r}_{ii}/|r_{ii}| & \text{if } r_{ii} \neq 0, \\ 1 & \text{if } r_{ii} = 0, \end{cases}$$

where r_{ii} is the i th diagonal element of R . It is straightforward to check that Λ is orthonormal, and $R' = \Lambda R$ is upper triangular with only nonnegative entries on its diagonal. Therefore, $Q' = Q\Lambda^H$ is orthonormal, and $A = Q'R'$ is a QR decomposition of A such that every diagonal element of R is nonnegative. Thus, throughout this section and the next, we assume the convention that the QR decomposition has only nonnegative diagonals. If A is nonsingular, the QR decomposition may be assumed to have only positive diagonals.

Proposition 13.4.1. *If $A \in M_n(\mathbb{F})$ is nonsingular, then the QR decomposition with only positive diagonals is unique.*

Proof. Let $A = Q_1 R_1$ and $A = Q_2 R_2$, where Q_1 and Q_2 are orthonormal and R_1 and R_2 are upper triangular with all positive elements on the diagonals. Note that

$$A^H A = R_1^H R_1 = R_2^H R_2,$$

which implies that

$$R_1 R_2^{-1} = R_1^{-H} R_2^H = (R_2 R_1^{-1})^H.$$

Since $R_1 R_2^{-1}$ is upper triangular and $(R_2 R_1^{-1})^H$ is lower triangular, it follows that both are diagonal. Writing the diagonal elements of R_1 and R_2 , respectively, as $R_1 = \text{diag}(a_1, a_2, \dots, a_n)$ and $R_2 = \text{diag}(b_1, b_2, \dots, b_n)$, it follows for each i that

$$\frac{a_i}{b_i} = \frac{b_i}{a_i},$$

which implies that each $a_i^2 = b_i^2$, and since each a_i and b_i are positive, we must have $a_i = b_i$ for each i . Thus, $R_1 R_2^{-1} = I$, which implies $R_1 = R_2$. Since $Q_1 = A R_1^{-1}$ and $Q_2 = A R_2^{-1}$, it follows that $Q_1 = Q_2$. $\square$

13.4.2 Iteration with the QR decomposition

Using the QR decomposition in an iterative fashion, we compute a full set of eigenvalues for a certain subclass of matrices $A \in M_n(\mathbb{F})$. We begin by setting $A_0 = A$ and computing the QR decomposition $A_0 = Q_0 R_0$. We then set $A_1 = R_0 Q_0$ and repeat. In other words, for each $k \in \mathbb{N}$, we compute the QR decomposition of

$$A_k = Q_k R_k \tag{13.26}$$

and then set

$$A_{k+1} = R_k Q_k. \tag{13.27}$$

We then show that in the limit as $k \rightarrow \infty$, the subdiagonals converge to zero and the diagonals converge to the eigenvalues of A .

Lemma 13.4.2. *Each A_k is similar to A .*

Proof. The proof is Exercise 13.16. $\square$

Lemma 13.4.3. *For each $k \in \mathbb{N}$ define*

$$U_k = Q_0 Q_1 \cdots Q_k \quad \text{and} \quad T_k = R_k \cdots R_1 R_0. \quad (13.28)$$

We have

$$A^{k+1} = U_k T_k \quad \text{and} \quad A_{k+1} = U_k^H A U_k. \quad (13.29)$$

Moreover, if A is nonsingular, then so is each T_k , and thus

$$A_{k+1} = T_k A T_k^{-1}. \quad (13.30)$$

Proof. We proceed by induction on $k \in \mathbb{N}$. It is easy to show that (13.29) and (13.30) hold for $k = 0$. Assume by the inductive hypothesis that $A^k = U_k T_k$ and $A_k = U_{k-1}^H A U_{k-1}$. If A is nonsingular, then we also assume $A_{k+1} = T_k A T_k^{-1}$. We have that

$$A^{k+2} = A A^{k+1} = A U_k T_k = U_k A^{k+1} T_k = U_k Q_{k+1} R_{k+1} T_k = U_{k+1} T_{k+1}$$

and

$$U_{k+1}^H A U_{k+1} = Q_{k+1}^H U_k^H A U_k Q_{k+1} = Q_{k+1}^H A_{k+1} Q_{k+1} = R_{k+1} Q_{k+1} = A_{k+2}.$$

Finally if A is nonsingular, then

$$T_{k+1} A T_{k+1}^{-1} = R_{k+1} T_k A T_k^{-1} R_{k+1}^{-1} = R_{k+1} A_{k+1} R_{k+1}^{-1} = R_{k+1} Q_{k+1} = A_{k+2}.$$

Thus, (13.29) and (13.30) hold for all $k \in \mathbb{N}$. $\square$

Theorem 13.4.4. *If $A \in M_n(\mathbb{F})$ is nonsingular and can be written as $A = P D P^{-1}$, where $D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is spectrally separated, that is, it satisfies*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0, \quad (13.31)$$

then the matrix $A_k = [a_{ij}^{(k)}]$, as defined in (13.27), satisfies the following:

- (i) *If $i > j$, then $\lim_{k \rightarrow \infty} a_{ij}^{(k)} = 0$.*
- (ii) *For each i , we have $\lim_{k \rightarrow \infty} a_{ii}^{(k)} = \lambda_i$.*

Proof. Let $P = QR$ be the QR decomposition of P with positive diagonals, and let $P^{-1} = LU$ be the LU decomposition of P^{-1} . Since

$$A = P D P^{-1} = Q R D R^{-1} Q^H,$$

we have that

$$Q^H A Q = R D R^{-1}$$

is upper triangular with diagonals $\lambda_1, \lambda_2, \dots, \lambda_n$ ordered in moduli. Note

$$A^k = P D^k P^{-1} = Q R D^k L U = Q R D^k L D^{-k} D^k U.$$

Since the diagonals of L are all one, we have

$$(D^k L D^{-k})_{ij} = \left(\frac{\lambda_i}{\lambda_j} \right)^k = \begin{cases} 0, & i < j, \\ 1, & i = j, \\ \rightarrow 0, & i > j \text{ as } k \rightarrow \infty, \end{cases} \quad (13.32)$$

or, in other words, $D^k L D^{-k} = I + E_k$, where $E_k \rightarrow 0$. Hence,

$$A^k = Q R D^k L D^{-k} D^k U = Q R (I + E_k) D^k U = Q (I + R E_k R^{-1}) R D^k U.$$

Writing $I + R E_k R^{-1} = \tilde{Q}_k \tilde{R}_k$ as the unique QR decomposition, where $\tilde{Q}_k \rightarrow I$ and $\tilde{R}_k \rightarrow I$, we have

$$A^k = (Q \tilde{Q}_k) (\tilde{R}_k R D^k U),$$

which is the product of an orthonormal matrix in an upper-triangular matrix. But we do not know that all of the diagonals of $\tilde{R}_k R D^k U$ are positive. Thus, for each k we choose the unitary diagonal matrix Λ_k so that $\Lambda_k \tilde{R}_k R D^k U$ has only positive diagonals. Hence,

$$\begin{aligned} T_k &= \Lambda_k \tilde{R}_k R D^k U \rightarrow \Lambda_k R D^k U, \\ U_k &= Q \tilde{Q}_k \Lambda_k^{-1} \rightarrow Q \Lambda_k^{-1}, \end{aligned}$$

which holds by uniqueness of the QR decomposition. It follows that

$$\Lambda_k^H A_{k+1} \Lambda_k = \Lambda_k^H U_k^H A U_k \Lambda_k \rightarrow Q^H A Q = R P^{-1} A P R^{-1} = R D R^{-1},$$

which is upper triangular. Thus, the lower-triangular values of A_{k+1} converge to zero and the diagonals of A_{k+1} converge to the diagonals of $R D R^{-1}$, which are the eigenvalues of A . $\square$

Example 13.4.5. Let

$$A = \frac{1}{3} \begin{bmatrix} 5 & 2 \\ 1 & 4 \end{bmatrix}.$$

By computing the QR iteration described above, we find that

$$\lim_{k \rightarrow \infty} A_k = \begin{bmatrix} 2 & 1/3 \\ 0 & 1 \end{bmatrix}.$$

From this, we see that $\sigma(A) = \{1, 2\}$. In this special case the upper-triangular portion converges, but this doesn't always happen.

Remark 13.4.6. The upper-triangular elements of A_{k+1} need not converge. In fact, it is not uncommon for them to oscillate in perpetuity as we iterate in k . As a result, we can't say that A_{k+1} converges, rather we only assert that the lower-triangular and diagonal portions do always converge as expected.

Example 13.4.7. Let

$$A = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}.$$

By computing the QR iteration described above, we find that

$$A_k \rightarrow \begin{bmatrix} -2 & (-1)^k \\ 0 & 1 \end{bmatrix}.$$

From this, we see that $\sigma(A) = \{-2, 1\}$. Note that the upper-triangular entry does not converge and in fact oscillates between positive and negative one.

Remark 13.4.8. If the hypothesis (13.31) is weakened, then QR iteration produces a block-upper-triangular matrix (similar to the original), where each block diagonal corresponds to eigenvalues of equal moduli. This follows because (13.32) generalizes to

$$|(D^k L D^{-k})_{ij}| = \left| \frac{\lambda_i}{\lambda_j} \right|^k = \begin{cases} 0, & |\lambda_i| < |\lambda_j|, \\ 1, & |\lambda_i| = |\lambda_j|, \\ \rightarrow 0, & |\lambda_i| > |\lambda_j| \text{ as } k \rightarrow \infty. \end{cases}$$

It is also important to note that the block diagonals will often not converge, but rather oscillate, since the convergence above is in modulus. Thus, we can perform QR iteration and then inspect A_k once k is sufficiently large that the lower-diagonal blocks converge to zero, and then we can analyze the individual blocks to find the spectrum. In the example below we consider a simple case where we can easily determine the spectrum of the oscillating block.

Example 13.4.9. Let

$$A = \begin{bmatrix} 5 & 1 & -5 & 0 \\ -4 & 12 & -4 & -3 \\ 10 & -8 & -4 & 3 \\ -2 & 13 & -8 & -2 \end{bmatrix}.$$

We can show that the eigenvalues of A are $\sigma(A) = \{2, 1, 4 + 3i, 4 - 3i\}$. By computing the QR iteration described above, we find that

$$A_{10000} = \begin{bmatrix} 9.8035 & 3.6791 & 1.6490 & -18.3022 \\ -11.6006 & -1.8035 & 8.5913 & 0.7464 \\ 0 & 0 & 2.0000 & -1.2127 \\ 0 & 0 & 0 & 1.0000 \end{bmatrix}.$$

Thus we can see instantly that two of the eigenvalues are 2 and 1, but we have to decompose the first block to get the other two eigenvalues. Since A_{10000} is block upper triangular, the eigenvalues of each block are the eigenvalues of the matrix (see Exercise 2.50), which is similar to A . Note that the block in question is

$$B = \begin{bmatrix} 9.8035 & 3.6791 \\ -11.6006 & -1.8035 \end{bmatrix},$$

which satisfies $\text{tr}(B) = 8$ and $\det(B) = 25$. From here it is easy to see that the eigenvalues of B are $4 \pm 3i$, and thus so are the eigenvalues of A_{10000} and A .

Remark 13.4.10. Anytime we have a real matrix with a conjugate pair of eigenvalues, which is often, the condition (13.31) is not satisfied. In this case, instead of the lower-triangular part of A_k converging, it will converge below the subdiagonal, and the 2×2 block corresponding to the conjugate pair will oscillate. One way to determine the eigenvalues is to compute the trace and determinant, as we did in the previous example (see Exercise 4.3). The challenge arises when there are many eigenvalues of the same modulus producing a large block. In the following section, we show how to remedy this problem.

Remark 13.4.11. In practice this isn't a very good method for computing the eigenvalues; it's too expensive computationally! The QR decomposition takes too many operations for this method to be practical for large matrices. The number of operations grows as a cubic polynomial in the dimension n of the matrix; thus, when n is large, doubling the matrix increases the number of operations by a factor of 8. And this is for each iteration! Since the number of QR iterations is likely to be much larger than n (especially if there are two eigenvalues whose moduli are close), the number of operations required by this method will have a leading order of more than n^4 . In the following section we show how to vastly improve the computation time and make QR iteration practical.

13.5 * Computing Eigenvalues II

The downside of the QR iteration described in the previous section is that it is computationally expensive and therefore not practical, unless the matrices are small. In this section we show how to substantially speed up QR iteration to make it much more practical for mid-sized matrices. For very large matrices, it may not be practical to compute all of the eigenvalues. Instead we may have to settle for just computing the largest eigenvalues. We finish this section with a discussion of a sort of a hybrid between the power method and QR iteration called the *Arnoldi algorithm*, which computes the largest eigenvalues.

The techniques in this section are based on three main observations. First, by using Householder transformations, we can transform a given matrix, via similarity, into a Hessenberg matrix, and then do QR iteration on the Hessenberg matrix instead. Second, if the initial matrix A_0 is Hessenberg, then so is each A_{k+1} in QR iteration. Finally, there's a very efficient way of doing the QR decomposition on Hessenberg matrices. By combining these three observations, we are able to do QR iteration relatively efficiently.

13.5.1 Preconditioning to Hessenberg Form

We begin by showing how to use Householder transformations to transform a matrix, via similarity, into a Hessenberg matrix. Recall from Section 3.4 that the Householder transform is an Hermitian orthonormal matrix that maps a given vector $\mathbf{x}$ to a multiple of $\mathbf{e}_1$ (see Lemma 3.4.7). This is used in Section 3.4 as a method for computing the QR decomposition, but we can also use it to transform a matrix into Hessenberg form.

Proposition 13.5.1. *Given $A \in M_n(\mathbb{F})$, there exists an orthonormal $Q \in M_n(\mathbb{F})$ such that $Q^H A Q$ is Hessenberg.*

Proof. We use a sequence of Householder transforms $H_1, H_2, \dots, H_{n-1}$ to eliminate the entries below the subdiagonals. For $k = 1$, write

$$A = \begin{bmatrix} a_{11} & \mathbf{a}_{12}^H \\ \mathbf{a}_{21} & A_{22} \end{bmatrix},$$

where $a_{11} \in \mathbb{F}$, where $\mathbf{a}_{12} \in \mathbb{F}^{n-1}$, and where $A_{22} \in M_{n-1}(\mathbb{F})$. Let $\hat{H}_1 \in M_{n-1}(\mathbb{F})$ denote the Householder matrix satisfying $\hat{H}_1 \mathbf{a}_{21} = v_1 \mathbf{e}_1 \in \mathbb{F}^{n-1}$, and let H_1 be the block-diagonal matrix $H_1 = \text{diag}(1, \hat{H}_1)$. Note that

$$H_1 A H_1 = \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \hat{H}_1 \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{a}_{12}^H \\ \mathbf{a}_{21} & A_{22} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^H \\ \mathbf{0} & \hat{H}_1 \end{bmatrix} = \begin{bmatrix} a_{11} & \mathbf{a}_{12}^H \hat{H}_1 \\ v_1 \mathbf{e}_1 & \hat{H}_1 A_{22} \hat{H}_1 \end{bmatrix}.$$

Therefore, all entries below the subdiagonal in the first column of $H_1 A H_1$ are zero.

For general $k > 1$, assume by the inductive hypothesis that we have

$$H_{k-1} \cdots H_1 A H_1 \cdots H_{k-1} = \begin{bmatrix} A_{11} & \mathbf{a}_{1,k} & A_{1,k+1} \\ 0 & \mathbf{a}_{k+1,k} & A_{k+1,k+1} \end{bmatrix}.$$

Here all the H_i are Hermitian and orthonormal, and $A_{11} \in M_{k \times k-1}(\mathbb{F})$ is upper Hessenberg. Also $\mathbf{a}_{k+1,k} \in \mathbb{F}^{n-k}$, $\mathbf{a}_{1,k} \in \mathbb{F}^k$, $A_{k+1,k+1} \in M_{k \times n-k}$, and $A_{k+1,k+1} \in M_{n-k}(\mathbb{F})$. Choose $\hat{H}_k \in M_{n-k}$ to be the Householder transformation such that $\hat{H}_k \mathbf{a}_{k+1,k} = v_k \mathbf{e}_1 \in \mathbb{F}^{n-k}$. Setting $H_k = \text{diag}(I_k, \hat{H}_k)$ gives

$$H_k H_{k-1} \cdots H_1 A H_1 \cdots H_{k-1} H_k = \begin{bmatrix} A_{11} & \mathbf{a}_{1,k} & A_{1,k+1} \hat{H}_k \\ \mathbf{0} & v_k \mathbf{e}_1 & \hat{H}_k A_{k+1,k+1} \hat{H}_k \end{bmatrix}.$$

Continuing for all $k < n$ shows that for $Q = H_1 H_2 \cdots H_{n-1}$ the matrix $Q^H A Q$ is upper Hessenberg. $\square$

13.5.2 QR Iteration for Hessenberg Matrices

Now we show that if the initial matrix A_0 in QR iteration is Hessenberg, then so is each A_{k+1} , where A_{k+1} is defined by (13.27).

Lemma 13.5.2. *The products AB and BA of a Hessenberg matrix $A \in M_n(\mathbb{F})$ and an upper-triangular matrix $B \in M_n(\mathbb{F})$ are Hessenberg.*

Proof. The matrix $A = [a_{ij}]$ satisfies $a_{ij} = 0$ whenever $i + 1 > j$, and the matrix $B = [b_{ij}]$ satisfies $b_{ij} = 0$ whenever $i > j$. The (i, k) entry of the product AB is given by

$$(AB)_{ik} = \sum_{j=1}^n a_{ij}b_{jk}. \quad (13.33)$$

Assume $i + 1 > k$ and consider $a_{ij}b_{jk}$. If $j < i + 1$, then $a_{ij} = 0$, which implies $a_{ij}b_{jk} = 0$. If $j \geq i + 1$, then since $j > k$, we have that $b_{jk} = 0$, which implies $a_{ij}b_{jk} = 0$. In other words, $a_{ij}b_{jk} = 0$ for all j , which implies that the sum (13.33) is zero. The proof that the product BA is Hessenberg is Exercise 13.19. $\square$

Theorem 13.5.3. *If A is Hessenberg, then the matrices Q_k and A_k given in (13.26) and (13.27) are also Hessenberg.*

Proof. The proof is Exercise 13.20. $\square$

13.5.3 QR Decomposition of Hessenberg Matrices

We now demonstrate an efficient way of doing the QR decomposition on Hessenberg matrices. Specifically, we show that the QR decomposition of a Hessenberg matrix $A \in M_n(\mathbb{F})$ can be performed by taking the product of $n-1$ rotation matrices (13.35) called *Givens rotations*. The reason this is useful is that the number of operations required to multiply by a Givens rotation is linear in n , and so the number of operations required to construct the QR decomposition of a Hessenberg matrix in this way is quadratic in n , whereas the number of operations required to construct the QR decomposition of a general matrix in $M_n(\mathbb{F})$ is cubic in n . Therefore, QR decomposition of Hessenberg matrices is much faster than QR decomposition of general matrices.

Definition 13.5.4. *Assume that $i < j$. A Givens rotation of θ radians between coordinates i and j is a matrix operator $G(i, j, \theta) \in M_n(\mathbb{F})$ of the form*

$$G(i, j, \theta) = I - (1 - \cos \theta)\mathbf{e}_i\mathbf{e}_i^T - \sin \theta\mathbf{e}_i\mathbf{e}_j^T + \sin \theta\mathbf{e}_j\mathbf{e}_i^T - (1 - \cos \theta)\mathbf{e}_j\mathbf{e}_j^T, \quad (13.34)$$

or, equivalently,

$$G(i, j, \theta) = \begin{bmatrix} I_{i-1} & 0 & 0 & 0 & 0 \\ 0 & \cos \theta & 0 & -\sin \theta & 0 \\ 0 & 0 & I_{j-i-1} & 0 & 0 \\ 0 & \sin \theta & 0 & \cos \theta & 0 \\ 0 & 0 & 0 & 0 & I_{n-j} \end{bmatrix}. \quad (13.35)$$

Remark 13.5.5. When a Givens rotation acts on a vector, it only modifies two components. Specifically, we have

$$G(i, j, \theta) \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_i \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ x_j \\ x_{j+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_i \cos \theta - x_j \sin \theta \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ x_i \sin \theta + x_j \cos \theta \\ x_{j+1} \\ \vdots \\ x_n \end{bmatrix}.$$

Similarly, left multiplying a matrix $A \in M_n(\mathbb{F})$ by a Givens rotation only modifies two rows of the matrix. The key point here is that the number of operations this multiplication takes is linear in the number of columns (n) of the matrix (roughly $10n$ operations). By contrast, multiplying two general $n \times n$ matrices together requires roughly $2n^3$ operations. Hence, when n is large, Givens rotations are much faster than general matrix multiplication.

Proposition 13.5.6. *Givens rotations satisfy*

- (i) $G(i, j, \theta)G(i, j, \phi) = G(i, j, \theta + \phi)$,
- (ii) $G(i, j, \theta)^\top = G(i, j, -\theta)$ (in particular, $G(i, j, \theta)^\top$ is also a Givens rotation),
- (iii) $G(i, j, \theta)$ is orthonormal.

Proof.

- (i) This is Exercise 13.21.
- (ii) This follows by taking the transpose of (13.35) and realizing that the cosine is an even function and the sine is an odd function.
- (iii) We have $G(i, j, \theta)G(i, j, \theta)^\top = G(i, j, \theta)G(i, j, -\theta) = G(i, j, 0) = I$. $\square$

Theorem 13.5.7. *Suppose $A \in M_n(\mathbb{F})$ is Hessenberg. Let $H_1 = A$ and define*

$$H_{k+1} = G_k^\top H_k, \quad k = 1, 2, \dots, n-1,$$

where $G_k = G(k, k+1, \theta_k)$, with

$$\theta_k = \text{Arctan} \left(\frac{h_{k+1,k}^{(k)}}{h_{k,k}^{(k)}} \right). \quad (13.36)$$

The matrix H_n is upper triangular and satisfies

$$A = G_1 G_2 \cdots G_{n-1} H_n.$$

In other words, the QR decomposition $A = QR$ of A is given by $Q = G_1 G_2 \cdots G_{n-1}$ and $R = H_n$.

Proof. We prove by induction that each H_k is Hessenberg, where the first $k - 1$ subdiagonal elements are zero. The case $k = 1$ is trivial. Assume $H_k = [h_{ij}^{(k)}]$ is Hessenberg and satisfies $h_{i+1,i}^{(k)} = 0$ for each $i = 1, 2, \dots, k-1$. By Exercise 13.22 the matrix H_{k+1} is Hessenberg. Moreover, Exercise 13.23 guarantees the subdiagonal entries $h_{i+1,i}^{(k+1)}$ all vanish for $i = 1, 2, \dots, k-1$. We complete the proof by showing that $h_{k+1,k}^{(k+1)} = 0$. We have

$$\begin{aligned} h_{k+1,k}^{(k+1)} &= \mathbf{e}_{k+1}^T G_k^T H_k \mathbf{e}_k \\ &= (\mathbf{e}_{k+1}^T - \sin \theta_k \mathbf{e}_k^T - (1 - \cos \theta_k) \mathbf{e}_{k+1}^T) H_k \mathbf{e}_k \\ &= -\mathbf{e}_k^T H_k \mathbf{e}_k \sin \theta_k + \mathbf{e}_{k+1}^T H_k \mathbf{e}_k \cos \theta_k \\ &= -h_{k,k}^{(k)} \sin \theta_k + h_{k+1,k}^{(k)} \cos \theta_k \\ &= 0, \end{aligned}$$

where the last step follows from (13.36). $\square$

Remark 13.5.8. As observed above, the number of operations required to carry out the QR decomposition of a Hessenberg matrix is quadratic in n , while the number required to construct the QR decomposition of a general matrix is cubic in n . This efficiency is why for QR iteration it's usually best to transform a matrix into Hessenberg form first and then do all the subsequent QR decompositions with Givens rotations.

13.5.4 The Arnoldi Method for Computing Eigenvalues

We finish this section by discussing the Arnoldi method, which approximates the largest eigenvalues of $A \in M_n(\mathbb{F})$ by performing QR iteration on the Arnoldi iterate H_k (see Section 13.3.1) instead of A . When $k = n$ the matrix H_n is similar to A , and therefore the eigenvalues of H_n are exactly the same as those of A (that is, within the accuracy of floating-point arithmetic). But for intermediate values of k , the eigenvalues of H_k are usually good approximations for the largest k eigenvalues (in modulus) of A .

Arnoldi iteration produces a $k \times k$ Hessenberg matrix H_k at each step, and the k eigenvalues of H_k are called the *Ritz eigenvalues* of A of A . There does not yet seem to be a good theory of how the Ritz eigenvalues converge to the actual eigenvalues. Nevertheless, in practice the Ritz eigenvalues do seem to converge (sometimes geometrically) to some of the eigenvalues of A . The Arnoldi method provides better approximations of the eigenvalues of A when the eigenspaces are orthogonal or nearly orthogonal, whereas when the eigenspaces are far from orthogonal, that is, nearly parallel, then the Ritz eigenvalues usually are worse approximations to the true eigenvalues of A . For more details on this, see [TB97].

Example 13.5.9. Here we use the Arnoldi method to calculate the Ritz eigenvalues of the matrix in Example 13.2.11. The square Hessenberg matrices (H_k) and the Ritz eigenvalues corresponding to Arnoldi step k are as follows:

H_k	Ritz Eigenvalues
$H_1 = [8.75]$	8.75
$H_2 = \begin{bmatrix} 8.75 & -5.472 \\ 5.403 & -1.495 \end{bmatrix}$	$3.627 \pm i1.823$
$H_3 = \begin{bmatrix} 8.75 & -5.472 & 2.230 \\ 5.403 & -1.495 & -0.134 \\ 0 & 0.724 & 0.672 \end{bmatrix}$	$3.416, 2.555 \pm i0.977$
$H_4 = \begin{bmatrix} 8.75 & -5.472 & 2.230 & -1.331 \\ 5.403 & -1.495 & -0.134 & 0.301 \\ 0 & 0.724 & 0.672 & -0.253 \\ 0 & 0 & 0.163 & 1.074 \end{bmatrix}$	$3.000, 2.000, 2.000, 2.000$

Remark 13.5.10. When $A \in M_n(\mathbb{F})$ is Hermitian, the Hessenberg matrices formed by Arnoldi iteration are tridiagonal and symmetric. In this case, the Arnoldi method is better understood theoretically, and is actually called the *Lanczos method*. In addition to the fact that A has orthogonal eigenspaces, the implementations are able to store fewer numbers by taking advantage of the symmetry of A ; and this greatly reduces the computation time.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed

the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

-
- 13.1. Let $A \in M_n(\mathbb{F})$ have components $A = [a_{ij}]$. For each i , let $R_i = \sum_{j \neq i} |a_{ij}|$ and let $D(a_{ii}, R_i)$ be the closed disk centered at a_{ii} with radius R_i . Prove that every eigenvalue λ of A lies within at least one of the disks $D(a_{ii}, R_i)$. Hint: Since the eigenvalue equation can be written as $\sum_{j=1}^n a_{ij}x_j = \lambda x_i$, split up the sum so that $\sum_{j \neq i} a_{ij}x_j = (\lambda - a_{ii})x_i$.
- 13.2. Assume $A \in M_n(\mathbb{F})$ is strictly diagonally dominant. Prove A is nonsingular. Hint: Use the previous exercise.
- 13.3. A matrix $C \in M_n(\mathbb{F})$ is called an *approximate inverse* of A if $r(I - CA) < 1$. Show that if C is an approximate inverse of A , then both A and C are invertible, and for any $\mathbf{x}_0 \in \mathbb{F}^n$ and $\mathbf{b} \in \mathbb{F}^n$, the map

$$\mathbf{x}_{n+1} = \mathbf{x}_n + C(\mathbf{b} - A\mathbf{x}_n)$$

converges to $\mathbf{x} = A^{-1}\mathbf{b}$.

- 13.4. Assume that $0 < a < b$. If $\varepsilon \geq 0$, prove that

$$\frac{a}{b} \leq \frac{a + \varepsilon}{b + \varepsilon} < 1.$$

Use this to prove that $\|B_{\text{GS}}\|_\infty \leq \|B_{\text{Jac}}\|_\infty$.

- 13.5. Given a matrix of the form

$$A = \begin{bmatrix} D_1 & B \\ C & D_2 \end{bmatrix},$$

where D_1 and D_2 are diagonal and invertible.

- (i) Prove that

$$B_{\text{Jac}} = \begin{bmatrix} 0 & -D_1^{-1}B \\ -D_2^{-1}C & 0 \end{bmatrix} \quad \text{and} \quad B_{\text{GS}} = \begin{bmatrix} 0 & -D_1^{-1}B \\ 0 & D_2^{-1}CD_1^{-1}B \end{bmatrix}.$$

- (ii) Prove that if $\lambda \in \sigma(B_{\text{Jac}})$, then $\lambda^2 \in \sigma(B_{\text{GS}})$.

- (iii) Conclude that $r(B_{\text{GS}}) = r(B_{\text{Jac}})^2$.

-
- 13.6. Prove Proposition 13.2.2. Hint: The Cayley-Hamilton theorem guarantees the existence of an annihilating polynomial. Use the well-ordering principle to conclude that one of least order exists. By normalizing the polynomial, you can make it monic. If there are two such minimal polynomials, then take their difference.
- 13.7. Prove Proposition 13.2.5. Hint: Use the spectral decomposition of A and the two lemmata preceding the statement of the proposition.

- 13.8. As an alternative proof that the matrix in (13.11) is nonsingular, show that all of its eigenvalues are one. Hint: Use the spectral mapping theorem (see Theorem 12.7.1) and the fact that all of the eigenvalues of a nilpotent matrix are zero (see Exercise 4.1).
- 13.9. Assuming $A \in M_n(\mathbb{F})$ is invertible, prove (13.9).

- 13.10. Use GMRES to solve the linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 2 \\ -4 \\ 1 \end{bmatrix}.$$

- 13.11. Let

$$A = \begin{bmatrix} 1 & 0 & c_1 \\ 0 & 1 & c_2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Prove for any $\mathbf{x}_0$ and $\mathbf{b}$ that GMRES converges to the exact solution after two steps.

- 13.12. Prove that if $\mathbf{r}_k$ is the k th residual in the GMRES algorithm for the system $A\mathbf{x} = \mathbf{b}$, then there exists a polynomial $q \in \mathbb{F}[z]$ of degree no more than k such that $\mathbf{r}_k = q(A)\mathbf{b}$.
- 13.13. Prove that if $A = UHU^{-1}$, where H is square and properly Hessenberg (meaning that the subdiagonal has no zero entries), and where $U \in M_n(F)$ has columns $\mathbf{u}_1, \dots, \mathbf{u}_n$, then $\text{span } \mathbf{u}_1, \dots, \mathbf{u}_j = \mathcal{K}_j(A, \mathbf{u}_1)$ for any $j \in \{1, \dots, n\}$.

- 13.14. Find the eigenvalues of the matrix

$$\begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 3 \\ 0 & 3 & 4 \end{bmatrix}$$

by using QR iteration.

- 13.15. What happens when QR iteration is applied to an orthonormal matrix? How does this relate to Theorem 13.4.4?
- 13.16. Prove Lemma 13.4.2.
- 13.17. One way to speed convergence of QR iteration is to use *shifting*. Instead of factoring $A_k = Q_k R_k$ and then setting $A_{k+1} = R_k Q_k$, factor $A_k - \sigma_k I$, where the shift σ_k is close to an eigenvalue of A . Show that for any $\sigma \in \mathbb{F}$, if $QR = A = \sigma I$ is the QR decomposition of $A - \sigma I$, then $RQ + \sigma I$ is orthonormally similar to A .

- 13.18. Put the following matrix into Hessenberg form:

$$\begin{bmatrix} 2 & 2 & 0 \\ 1 & 4 & 1 \\ 3 & 1 & 0 \end{bmatrix}$$

- 13.19. Prove that the product BA of a Hessenberg matrix $A \in M_n(\mathbb{F})$ and an upper triangular matrix $B \in M_n(\mathbb{F})$ is a Hessenberg matrix.
- 13.20. Prove Theorem 13.5.3.
- 13.21. Prove that Givens rotations $G(i, j, \theta)$ satisfy the identity $G(i, j, \theta)G(i, j, \phi) = G(i, j, \theta + \phi)$.
- 13.22. Prove that H_{k+1} as defined in Theorem 13.5.7 is Hessenberg.
- 13.23. In the proof of Theorem 13.5.7, show that $h_{i+1,i}^{(k+1)} = 0$ for $i = 1, 2, \dots, k-1$.
-

Notes

Our treatment of Krylov methods is based in part on [TB97] and [IM98]. Exercise 13.15 is from [TB97].

For details on the stability and computational complexity of the methods described in this chapter, see [GVL13, CF13].

14

Spectra and
Pseudospectra

I am not strange. I am just not normal.
—Salvador Dalí

Recall from Section 7.5.4 that the eigenvalue problem can be ill conditioned; that is, a very small change in a matrix can produce a relatively large change in its eigenvalues. This happens when the eigenspaces are nearly parallel. By contrast we saw in Corollary 7.5.15 that the eigenvalue problem for normal matrices is well conditioned. In other words, when matrices have orthogonal or nearly orthogonal eigenspaces, the eigenvalue problem is well conditioned, and when the eigenspaces are far from orthogonal, that is, very nearly parallel, the eigenvalue problem is ill conditioned.

When a problem is ill conditioned, two nearly indistinguishable inputs can have very different outputs, thus calling into question the reliability of the solution in the presence of any kind of uncertainty, including that arising from finite-precision arithmetic. For example, if two nearly indistinguishable matrices have very different eigenvalues and therefore have wildly different behaviors, then the results of most computations involving those matrices probably cannot be trusted.

An important example of this occurs with the iterative methods described in Chapter 13 for solving linear systems. In Section 13.1 we described three iterative methods for solving linear systems by taking powers of matrices. If the eigenvalues of the matrices are contained in the open unit disk, then the method converges (see Theorem 13.1.1). However, even when all the eigenvalues are contained in the open unit disk, a problem can arise if one or more of the eigenvalues are very nearly unit length and the corresponding eigenspace is nearly parallel to another of its eigenspaces. In this case, it is possible that these matrices will be essentially indistinguishable, numerically, from those that have eigenvalues larger than one. Because of ill conditioning, these iterative methods can fail to converge in practice, even when they satisfy the necessary and sufficient conditions for convergence. Moreover, even when the iterative methods do converge, ill conditioning can drastically slow their convergence.

This chapter is about *pseudospectral theory*, which gives tools for analyzing how conditioning impacts results and methods that depend on the eigenvalues of

matrices. In the first section we define the pseudospectrum and provide a few equivalent definitions. One of these definitions describes the pseudospectrum in terms of the spectra of nearby operators, which gives us a framework for connecting convergence to conditioning. Another equivalent definition uses the resolvent. Recall that the poles of the resolvent correspond to the eigenvalues. The pseudospectrum corresponds to the regions of the complex plane where the norm of the resolvent is large but not necessarily infinite, indicating eigenvalues are nearby. These regions give a lot of information about the behavior of these matrices in computations.

In the second section, we discuss the transient behavior of matrix powers. Consider the sequence $(\|A^k\|)_{k \in \mathbb{N}}$, generated by a matrix $A \in M_n(\mathbb{C})$. The sequence converges to zero as $k \rightarrow \infty$ if and only if the spectral radius $r(A)$ of A is less than one, but before it goes to zero, it may actually grow first. And it need not go to zero very quickly. Since convergence of many iterative methods depends upon these matrix powers approaching zero, it becomes important to understand not just what happens for large values of k , but also what happens for small and intermediate values of k . When does the sequence converge monotonically? And when does it grow first before decaying?

In the second section we address these questions and provide upper and lower bounds for the sequence of powers via the Kreiss matrix theorem. We also discuss preconditioning, which is a way of transforming the matrices used in iterative methods into new matrices that have better behaved sequences with faster convergence. These are useful when dealing with a poorly conditioned matrix with eigenvalues close to the unit circle. The pseudospectrum gives insights into how to choose a good preconditioner.

In the final section we prove the Kreiss matrix theorem by appealing to a key lemma by Spijker. The proof of Spijker's lemma is the longest part of this proof.

14.1 The Pseudospectrum

In this section, we define the pseudospectrum and describe some of its properties as well as several alternative definitions. We consider the role that conditioning plays in describing the pseudospectrum and then present a few results that provide upper and lower bounds on the pseudospectrum.

Remark 14.1.1. The definition of the pseudospectrum and most of the results of this chapter extend very naturally to infinite-dimensional spaces and general operator norms, but in order to avoid some technical complications that would take us too far afield, we limit ourselves to $\mathbb{C}^n$ and to the standard 2-norm on matrices.

14.1.1 Definition and Basic Properties

The pseudospectrum of a matrix $A \in M_n(\mathbb{C})$ is just the locus of all eigenvalues of matrices sufficiently near to A .

Definition 14.1.2. Let $A \in M_n(\mathbb{C})$ and $\varepsilon > 0$. The ε -pseudospectrum of A is the set

$$\sigma_\varepsilon(A) = \{z \in \mathbb{C} \mid z \in \sigma(A + E) \text{ for some } E \in M_n(\mathbb{C}) \text{ with } \|E\| < \varepsilon\}. \quad (14.1)$$

The elements $\sigma_\varepsilon(A)$ are called the ε -pseudoeigenvalues of A .

In some sense the pseudospectra represent the possible eigenvalues of the matrix when you throw in a little dirt. In mathematics we call these *perturbations*. Ideally when we solve problems, we want the solutions to be robust to small perturbations. Indeed, if your answer depends on infinite precision and no error, then it probably has no connection to real-world phenomena and is therefore of little use.

Remark 14.1.3. We always have $\sigma(A) \subset \sigma_\varepsilon(A)$ for all $\varepsilon > 0$, since $E = 0$ trivially satisfies $\|E\| < \varepsilon$.

Remark 14.1.4. The following two properties of pseudospectra follow immediately from the definition:

(i) If $0 < \varepsilon_1 \leq \varepsilon_2$, then $\sigma_{\varepsilon_1}(A) \subset \sigma_{\varepsilon_2}(A)$.

(ii) $\bigcap_{\varepsilon > 0} \sigma_\varepsilon(A) = \sigma(A)$.

By combining these two facts, we can identify $\sigma(A)$ as the $\varepsilon = 0$ pseudospectrum.

Example 14.1.5. Recall the nonnormal, ill-conditioned matrix A of Example 7.5.16,

$$A = \begin{bmatrix} 1 & 1000 \\ 0.001 & 1 \end{bmatrix},$$

which has eigenvalues $\{0, 2\}$. In Example 7.5.16 we saw that if A was perturbed by a matrix

$$E = \begin{bmatrix} 0 & 0 \\ -0.001 & 0 \end{bmatrix},$$

the eigenvalues of $A + E$ became a double eigenvalue at 1. Figure 14.1(a) depicts the various ε -pseudospectra of A for $\varepsilon = 10^{-2}$, $10^{-2.5}$, and 10^{-3} . That is, the figure depicts eigenvalues of matrices of the form $A + E$, where E is a matrix with 2-norm ε .

14.1.2 Equivalent Definitions

There are a few equivalent alternative definitions of the pseudospectrum.

Theorem 14.1.6. *Given $A \in M_n(\mathbb{C})$ and $\varepsilon > 0$, the following sets are equal:*

(i) *The ε -pseudospectrum $\sigma_\varepsilon(A)$ of A .*

(ii) *The set of $z \in \mathbb{C}$ such that*

$$\|(zI - A)\mathbf{v}\| < \varepsilon \tag{14.2}$$

for some $\mathbf{v} \in \mathbb{F}^n$ with $\|\mathbf{v}\| = 1$. The vectors $\mathbf{v}$ are the ε -pseudoeigenvectors of A corresponding to the ε -pseudoeigenvalues $z \in \mathbb{C}$.

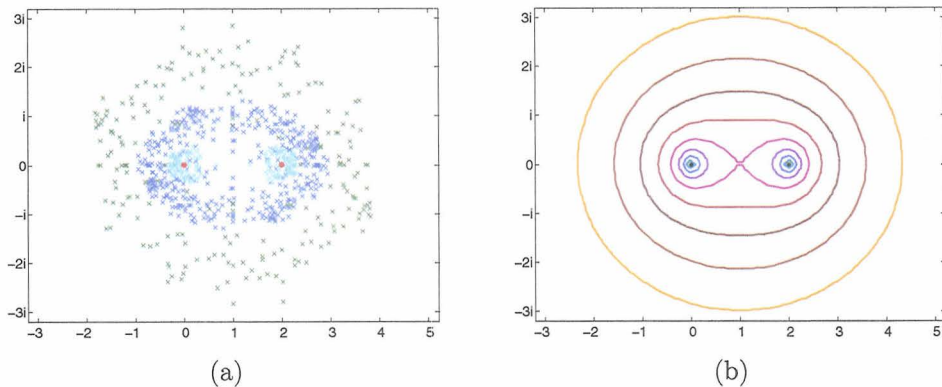


Figure 14.1. Two depictions of pseudospectra for the matrix A of Examples 7.5.16 and 14.1.5. In panel (a), the two eigenvalues of A are plotted in red, along with plots of eigenvalues of randomly chosen matrices of the form $A + E$ with $\|E\|_2 = 10^{-3}$ (cyan), $10^{-2.5}$ (blue), and 10^{-2} (green). Panel (b) shows various contour lines for the resolvent. By Theorem 14.1.6(iii), these contour lines represent the boundaries of various ε -pseudospectra.

(iii) The set of $z \in \mathbb{C}$ such that

$$\|R(A, z)\| > \varepsilon^{-1}, \quad (14.3)$$

where $R(A, z)$ is the resolvent of A evaluated at z .

Proof. We establish equivalence by showing that each set is contained in the other.

(i) $\subset$ (ii). Suppose $(A + E)\mathbf{v} = z\mathbf{v}$ for some $\|E\| < \varepsilon$ and some unit vector $\mathbf{v}$. In this case, $\|(zI - A)\mathbf{v}\| = \|E\mathbf{v}\| \leq \|E\|\|\mathbf{v}\| < \varepsilon$.

(ii) $\subset$ (iii). Assume $(zI - A)\mathbf{v} = s\mathbf{u}$ for some unit vectors $\mathbf{v}, \mathbf{u} \in \mathbb{C}^n$ and $0 < s < \varepsilon$. In this case, $(zI - A)^{-1}\mathbf{u} = s^{-1}\mathbf{v}$, which implies $\|(zI - A)^{-1}\| \geq s^{-1} > \varepsilon^{-1}$.

(iii) $\subset$ (i). If $\|(zI - A)^{-1}\| > \varepsilon^{-1}$, then by definition of the norm of an operator, there exists a unit vector $\mathbf{u} \in \mathbb{C}^n$ such that $\|(zI - A)^{-1}\mathbf{u}\| > \varepsilon^{-1}$. Thus, there exists a unit vector $\mathbf{v} \in \mathbb{C}^n$ and a positive $s < \varepsilon$ such that $(zI - A)^{-1}\mathbf{u} = s^{-1}\mathbf{v}$. In this case, we have that $s\mathbf{u} = (zI - A)\mathbf{v}$, which implies that $(A + s\mathbf{u}\mathbf{v}^H)\mathbf{v} = z\mathbf{v}$. Setting $E = s\mathbf{u}\mathbf{v}^H$, it suffices to show that $\|E\| = s < \varepsilon$. But Exercise 4.31(ii) implies that

$$\|E\|^2 = s^2\|(\mathbf{v}\mathbf{u}^H)(\mathbf{u}\mathbf{v}^H)\| \leq s^2\|\mathbf{v}\mathbf{v}^H\| \leq s^2,$$

where the last inequality follows from the fact that the largest singular value of the matrix $\mathbf{v}\mathbf{v}^H$ is 1. Thus, z is an eigenvalue of $A + E$ for some $\|E\| < \varepsilon$. $\square$

Remark 14.1.7. Each of these equivalent definitions of the pseudospectrum has its advantages. The definition (i) is well motivated, but perhaps more difficult to visualize than the more traditional definition (iii) in terms of resolvents, which shows that $\sigma_\varepsilon(A)$ is the open subset of $\mathbb{C}$ bounded by the ε^{-1} level set of the normed resolvent.

Remark 14.1.8. Recall that the spectrum of an operator A is the locus where the resolvent $\|(zI - A)^{-1}\|$ is infinite. The fact that definitions (i) and (iii) are equivalent means that the locus where $\|(zI - A)^{-1}\|$ is large, but not necessarily infinite, gives information about how the spectrum will change when the operator is slightly perturbed.

Example 14.1.9. Let

$$A = \begin{bmatrix} 1+i & 0 & i \\ -i & 0.5 & 0 \\ 0.3i & 0.5 & 0.7 \end{bmatrix}.$$

Figure 14.2(a) shows a plot of the norm of the resolvent of A as a function of $z \in \mathbb{C}$. The poles occur at the eigenvalues of A , and the points in the plane where the plot is greater than ε^{-1} form the ε -pseudospectrum of A . Figure 14.2(b) shows the level curves of $\|R(A, z)\|$ for the matrix A , corresponding to the points where $\|R(A, z)\| = \varepsilon^{-1}$ for various choices of ε . For a given choice of ε , the interior of the region bounded by the curve $\|R(A, z)\| = \varepsilon^{-1}$ is the ε -pseudospectrum.

We can also give another equivalent form of the ε -pseudospectrum, but unlike the previous three definitions, this does not generalize to infinite-dimensional spaces or to norms other than the 2-norm.

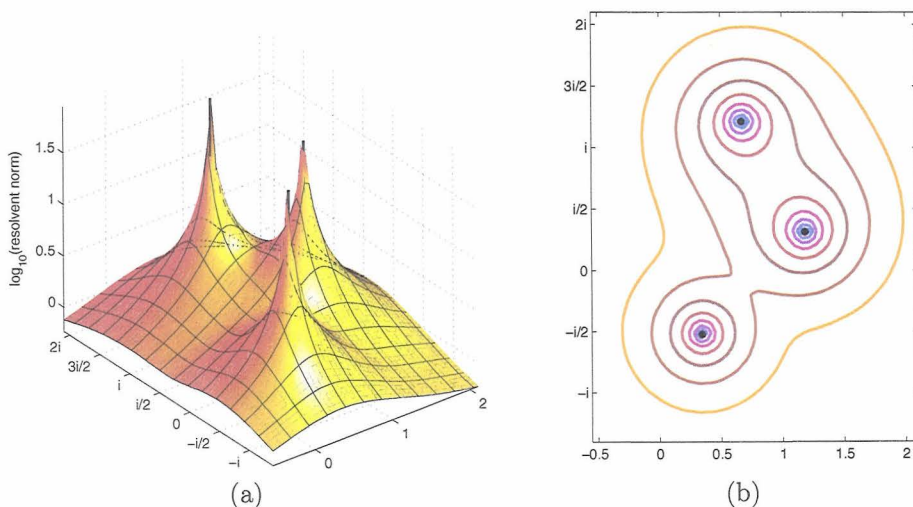


Figure 14.2. A plot of $\|R(A, z)\|$ for the matrix A in Example 14.1.9 as z varies over the complex plane. These are represented as (a) a three-dimensional plot and (b) a topographical map.

Proposition 14.1.10. *If $A \in M_n(\mathbb{C})$, the set $\sigma_\varepsilon(A)$ is equal to the set of $z \in \mathbb{C}$ such that*

$$s_{\min}(zI - A) < \varepsilon, \quad (14.4)$$

where $s_{\min}(zI - A)$ is the smallest singular value of $(zI - A)$.

Proof. The equivalence of (14.4) and (14.2) is Exercise 14.4. $\square$

Example 14.1.11. Consider the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3628800 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -10628640 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12753576 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -8409500 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 3416930 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -902055 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 157773 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -18150 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1320 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -55 \end{bmatrix} \in M_{10}(\mathbb{R}).$$

It is not hard to verify that the characteristic polynomial of A is the degree-ten Wilkinson polynomial $p(x) = \prod_{k=1}^{10}(x - k)$, and hence the spectrum is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

We can visualize the pseudospectra of A in two different ways. First, using the resolvent definition (iii) of the pseudospectrum, we can plot the boundaries of $\sigma_\varepsilon(A)$ for $\varepsilon = 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}$. These consist of tiny circles around the eigenvalues 1 and 2, and then some larger oblong curves around the larger eigenvalues, as depicted in Figure 14.3(a). Alternatively, using the perturbation definition (i) we can perturb A by various random matrices E with norm no more than 10^{-8} and plot the eigenvalues of the perturbed matrices in the complex plane. This is seen in Figure 14.3(b). Again, a perturbation of the matrix has almost no effect on the eigenvalues near 1 and 2, but has a large effect on the eigenvalues near 8, 9, and 10.

14.1.3 Pseudospectra and Conditioning

Since the ε -pseudospectrum of A is the set of all eigenvalues of all ε perturbations of A , it is very closely tied to the condition of the eigenvalue problem for A . Any bounds on the pseudospectrum correspond to bounds on how much the eigenvalues can move when the matrix is perturbed. In this section we prove the Bauer–Fike theorem (Corollary 14.1.15), which gives a useful bound on the pseudospectrum of a diagonalizable matrix and hence on the condition number of the eigenvalue problem for such matrices.

Proposition 14.1.12. *The following inequalities hold for all $A \in M_n(\mathbb{C})$ and $z \in \mathbb{C}$:*

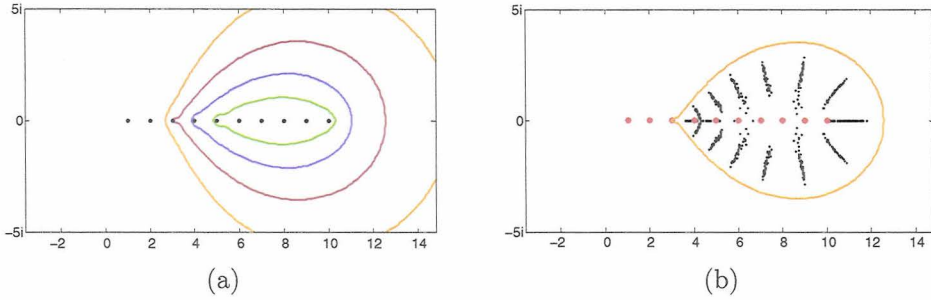


Figure 14.3. In (a), the eigenvalues (black) of the matrix A from Example 14.1.11, and the boundaries of $\sigma_\varepsilon(A)$. The boundary for $\varepsilon = 10^{-7}$ is the outer (yellow) curve, along with some invisibly tiny circles around the eigenvalues 1 and 2. The boundary for $\varepsilon = 10^{-10}$ is the inner (green) oval, along with invisibly tiny circles around 1, 2, 3, and 4. In (b), the eigenvalues (red) of A along with the boundary (yellow) of $\sigma_\varepsilon(A)$ for $\varepsilon = 10^{-8}$ and the eigenvalues (black) of various perturbed matrices of the form $A + E$, where E is chosen randomly with $\|E\|_2 < 10^{-8}$.

(i)

$$\|(zI - A)^{-1}\| \geq \frac{1}{\text{dist}(z, \sigma(A))}. \quad (14.5)$$

(ii) If A is diagonalizable and has the form $A = VDV^{-1}$, where D is diagonal, then

$$\|(zI - A)^{-1}\| \leq \frac{\kappa(V)}{\text{dist}(z, \sigma(A))}, \quad (14.6)$$

where $\kappa(V) = \|V\|\|V^{-1}\|$ is the condition number of V (see Section 7.5.3).

(iii) If A is normal, then

$$\|(zI - A)^{-1}\| = \frac{1}{\text{dist}(z, \sigma(A))}. \quad (14.7)$$

Proof. If $z \in \sigma(A)$, then both $\|(zI - A)^{-1}\| = \infty$ and $\text{dist}(z, \sigma(A)) = 0$ and all the relations hold trivially. Thus, we assume below that $z \notin \sigma(A)$.

(i) If $Av = \lambda v$ for some unit vector $v \in \mathbb{C}^n$, then $(zI - A)v = (z - \lambda)v$. Thus, $(zI - A)^{-1}v = (z - \lambda)^{-1}v$, which implies that

$$\frac{1}{|z - \lambda|} = \|(zI - A)^{-1}v\| \leq \|(zI - A)^{-1}\|.$$

Since this holds for all $\lambda \in \sigma(A)$, we have that

$$\|(zI - A)^{-1}\| \geq \max_{\lambda \in \sigma(A)} \frac{1}{\text{dist}(z, \lambda)} = \frac{1}{\text{dist}(z, \sigma(A))}.$$

- (ii) Since $\|(zI - A)^{-1}\| = \|V(zI - D)^{-1}V^{-1}\| \leq \|V\|\|V^{-1}\|\|(zI - D)^{-1}\|$, it follows that

$$\|(zI - A)^{-1}\| \leq \kappa(V)\|(zI - D)^{-1}\| \leq \kappa(V) \max_{\lambda \in \sigma(A)} \frac{1}{|z - \lambda|} = \frac{\kappa(V)}{\text{dist}(z, \sigma(A))}.$$

- (iii) If A is normal, then it is orthonormally diagonalizable as $A = VDV^H$. But $V^H V = I$ implies $\|V\| = \|V^H\| = 1$, and thus the condition number of V satisfies $\kappa(V) = 1$. The result now follows from (14.5) and (14.6). $\square$

Remark 14.1.13. The condition number $\kappa(V)$ depends on the choice for the diagonalizing transformation V , which is not unique. Indeed, if we rescale the column vectors of V (that is, the eigenvectors) with $V\Lambda$ where Λ is a nonsingular diagonal matrix, then $V\Lambda$ still diagonalizes A , that is, $(V\Lambda)D(V\Lambda)^{-1} = (V\Lambda)D\Lambda^{-1}V^{-1} = VDV^{-1} = A$, yet the condition number $\kappa(V\Lambda)$ is different and can be drastically so in some cases. In fact, by making the eigenvectors in V sufficiently large or small, we can make the condition number as large as we want to.

Remark 14.1.14. The number

$$\kappa_\sigma(A) = \inf\{\kappa(V) \mid A = VDV^{-1}\}$$

is sometimes called the *spectral condition number* of A . We may substitute κ_σ for $\kappa(V)$ in (14.6). We do not address this further here, but some references about bounds on the spectral condition number may be found in the notes at the end of this chapter.

Corollary 14.1.15 (Bauer–Fike Theorem). *If $A \in M_n(\mathbb{C})$ satisfies $A = VDV^{-1}$ with D diagonal, then for any $\varepsilon > 0$ the ε -pseudospectrum satisfies the inequality of sets*

$$\sigma(A) + B(0, \varepsilon) \subset \sigma_\varepsilon(A) \subset \sigma(A) + B(0, \varepsilon\kappa), \quad (14.8)$$

where $\kappa = \kappa(V)$ is the condition number for V . In the special case that A is normal, we have that

$$\sigma_\varepsilon(A) = \sigma(A) + B(0, \varepsilon) = \{z \in \mathbb{C} : |z - \lambda| < \varepsilon \text{ for at least one } \lambda \in \sigma(A)\}.$$

Remark 14.1.16. The Bauer–Fike theorem shows that the (absolute) condition number of the eigenvalue problem for a diagonalizable A is bounded by the condition number $\kappa(V)$ for every V diagonalizing A , and therefore it is bounded by the spectral condition number $\kappa_\sigma(A)$.

Remark 14.1.17. The Bauer–Fike theorem says that the pseudospectrum is not especially useful when A is normal because it can be constructed entirely from the information about the spectrum alone.

14.2 Asymptotic and Transient Behavior

In this section, we study the long- and short-term behavior of the norm of the powers of a matrix $A \in M_n(\mathbb{F})$, that is, $\|A^k\|$ for $k \in \mathbb{N}$. As discussed in the introduction of this chapter, this class of functions is widely used in applications. For example, many iterative methods of linear algebra, including all of those described in Section 13.1, depend on the convergence of this sequence of powers.

The long-term (*asymptotic*) behavior of $\|A^k\|$ is characterized by the spectrum of A . In particular, if the spectral radius $r(A)$ is less than 1, then $\|A^k\| \rightarrow 0$ as $k \rightarrow \infty$; if $r(A) > 1$, then $\|A^k\| \rightarrow \infty$ as $k \rightarrow \infty$; and if $r(A) = 1$, then $\|A^k\|$ may either be bounded or grow to infinity depending on the circumstances.

Even though the asymptotic behavior of $\|A^k\|$ depends exclusively on the eigenvalues of the underlying matrix, for intermediate values of k , the quantities can grow quite large before converging to zero. We call this intermediate stage the *transient* behavior.

In this section we show that the transient behavior depends on the pseudospectrum of $A \in M_n(\mathbb{F})$ instead of the spectrum. Moreover, the Kreiss matrix theorem gives both upper and lower bounds on the transient behavior. We state the Kreiss matrix theorem in this section and prove it in the next.

14.2.1 Asymptotic Behavior

The asymptotic behavior of the sequence $\|A^k\|$ can be summarized in the following theorem.

Theorem 14.2.1. *For any $A \in M_n(\mathbb{F})$,*

- (i) *if $r(A) < 1$, then $\|A^k\| \rightarrow 0$ as $k \rightarrow \infty$;*
- (ii) *if $r(A) > 1$, then $\|A^k\| \rightarrow \infty$ as $k \rightarrow \infty$;*
- (iii) *if $r(A) = 1$, then $\|A^k\|$ is bounded if and only if all the eigenvalues λ satisfying $|\lambda| = 1$ are semisimple, that is, they have no eigennilpotents.*

Proof.

- (i) This follows from Exercise 12.15.
- (ii) If $r(A) > 1$, then there exists $\lambda \in \sigma(A)$ such that $|\lambda| > 1$. Since $|\lambda|^k \leq \|A^k\|$ we have that $\|A^k\| \rightarrow \infty$ as $k \rightarrow \infty$.
- (iii) See Exercises 14.9 and 14.10. $\square$

14.2.2 Transient Behavior

The obvious analogue of the spectral radius for pseudospectra is the *pseudospectral radius*. The pseudospectral radius and a related quantity called the *Kreiss matrix constant* are important for understanding the transient behavior of $\|A^k\|$.

Definition 14.2.2. *For a matrix $A \in M_n(\mathbb{C})$ and any $\varepsilon > 0$, the ε -pseudospectral radius $r_\varepsilon(A)$ of A is given by*

$$r_\varepsilon(A) = \sup\{|z| : z \in \sigma_\varepsilon(A)\}.$$

The Kreiss constant is the quantity

$$K(A) = \sup_{\varepsilon > 0} \frac{r_\varepsilon(A) - 1}{\varepsilon}. \quad (14.9)$$

The Kreiss constant also measures how fast the norm of the resolvent blows up inside z , the open unit disk, as shown in the following equivalent, alternative definition of the Kreiss matrix constant.

Proposition 14.2.3. *If $A \in M_n(\mathbb{F})$, then*

$$K(A) = \sup_{|z| > 1} (|z| - 1) \|(zI - A)^{-1}\|. \quad (14.10)$$

Proof. Fix $\varepsilon > 0$. Choose z_0 so that $|z_0| = r_\varepsilon(A)$. By continuity of the resolvent and the norm, we have that $\|(z_0I - A)^{-1}\| \geq \varepsilon^{-1}$. Thus,

$$(|z_0| - 1) \|(z_0I - A)^{-1}\| \geq \frac{r_\varepsilon(A) - 1}{\varepsilon}.$$

Taking the supremum over $\mathbb{C}$ gives

$$\sup_{z \in \mathbb{C}} (|z| - 1) \|(zI - A)^{-1}\| \geq \frac{r_\varepsilon(A) - 1}{\varepsilon}.$$

Since the supremum will never occur when $|z| \leq 1$, we can restrict the domain and write

$$\sup_{|z| > 1} (|z| - 1) \|(zI - A)^{-1}\| \geq \frac{r_\varepsilon(A) - 1}{\varepsilon}.$$

Since this holds for all $\varepsilon > 0$, we have

$$\sup_{|z| > 1} (|z| - 1) \|(zI - A)^{-1}\| \geq \sup_{\varepsilon > 0} \frac{r_\varepsilon(A) - 1}{\varepsilon} = K(A).$$

To establish the other direction, fix $|z| > 1$. Define $\varepsilon^{-1} = \|(zI - A)^{-1}\|$. Thus, z is the boundary of σ_{ε_0} . Hence, $r_{\varepsilon_0}(A) \geq |z|$. This yields

$$\frac{r_{\varepsilon_0}(A) - 1}{\varepsilon_0} \geq \frac{|z| - 1}{\varepsilon} = (|z| - 1) \|(zI - A)^{-1}\|.$$

Taking the supremum over all $\varepsilon > 0$ gives

$$K(A) = \sup_{\varepsilon > 0} \frac{r_{\varepsilon_0}(A) - 1}{\varepsilon_0} \geq (|z| - 1) \|(zI - A)^{-1}\|.$$

Since this holds for all $|z| > 1$, we have

$$K(A) \geq \sup_{|z| > 1} (|z| - 1) \|(zI - A)^{-1}\|. \quad \square$$

Remark 14.2.4. The Kreiss constant is only useful when $r(A) \leq 1$ because $K(A) = \infty$ whenever $r(A) > 1$, as can be seen from (14.10).

Proposition 14.2.5. *If $A \in M_n(\mathbb{C})$ is normal and $r(A) \leq 1$, then $K(A) = 1$.*

Proof. Since A is normal, Exercise 14.6 shows that $r_\varepsilon(A) = r(A) + \varepsilon$. We have

$$K(A) = \sup_{\varepsilon > 0} \frac{r_\varepsilon(A) - 1}{\varepsilon} = \sup_{\varepsilon > 0} \frac{r(A) + \varepsilon - 1}{\varepsilon} \leq 1.$$

Moreover,

$$\lim_{\varepsilon \rightarrow \infty} \frac{r(A) + \varepsilon - 1}{\varepsilon} = 1,$$

which implies that $K(A) = 1$. $\square$

We are now prepared to give a lower bound on the transient behavior of $\|A^k\|$.

Lemma 14.2.6. *If $A \in M_n(\mathbb{F})$, then*

$$K(A) \leq \sup_k \|A^k\|. \quad (14.11)$$

Proof. Let $M = \sup_k \|A^k\|$. If $M = \infty$, the result follows trivially. We now assume M is finite, and thus $\|A\| \leq 1$. Choose $z \in \mathbb{C}$ such that $|z| > 1$. By Theorem 12.3.8, we can write

$$\|(zI - A)^{-1}\| = \left\| \sum_{k=0}^{\infty} \frac{A^k}{z^{k+1}} \right\| \leq \sum_{k=0}^{\infty} \frac{\|A^k\|}{|z|^{k+1}} \leq \frac{M}{|z|} \sum_{k=0}^{\infty} \left(\frac{1}{|z|} \right)^k = \frac{M}{|z| - 1}.$$

Thus,

$$(|z| - 1)\|(zI - A)^{-1}\| \leq M.$$

Since z was arbitrary, we may take the supremum of the left side to get the desired result. $\square$

We are now prepared to state the Kreiss matrix theorem. The proof is given in Section 14.3.

Theorem 14.2.7 (Kreiss Matrix Theorem). *If $A \in M_n(\mathbb{F})$, then*

$$K(A) \leq \sup_{k \in \mathbb{N}} \|A^k\| \leq e n K(A). \quad (14.12)$$

Remark 14.2.8. The Kreiss matrix theorem gives both upper and lower bounds on the transient behavior of $\|A^k\|$. When the Kreiss constant is greater than one, it means that the sequence $(\|A^k\|)_{k=0}^{\infty}$ grows before decaying back to zero. If the Kreiss constant is large, then the transient phase is nontrivial and it will take many iterations before A^k converges to zero. This means that iterative methods may take a while to converge. By contrast, if the Kreiss constant is close to one, then the transient phase should be relatively brief by comparison and convergence should be relatively fast.

Remark 14.2.9. The original statement of the Kreiss theorem did not actually look much like the theorem above. The original version of the right-hand inequality in (14.12), proven by Kreiss in 1962, was $\sup_{k \in \mathbb{N}} \|A^k\| \leq CK(A)$, where $C \sim c^{n^n}$. Over time the bound has been sharpened through a series of improvements.

The current bound is sharp in the following sense. Although there may not be matrices for which equality is actually attained, the inequality is the best possible in the sense that if $(\|A^k\|)_{k \in \mathbb{N}} \leq Cn^\alpha K(A)$ for all A , then α can be no smaller than one. Similarly the factor e is the best possible, since if $(\|A_\alpha^k\|)_{k \in \mathbb{N}} \leq Cn^\alpha K(A)$ and $\alpha = 1$, then C can be no smaller than e . For a good historical survey and a more complete treatment of the Kreiss matrix theorem, see [TE05].

Application 14.2.10 (Markov Chains and the Cutoff Phenomenon).

A finite-state stationary Markov chain can be described as an iterative system $\mathbf{x}_{k+1} = P\mathbf{x}_k$, where the matrix $P \in M_n(\mathbb{R})$ is nonnegative with each column summing to one and the initial vector $\mathbf{x}_0$ is also nonnegative with the components summing to one.

Assuming P is irreducible (see Definition 12.8.9), the power method (Theorem 12.7.8) implies the sequence $\mathbf{x}_k = P^k \mathbf{x}_0$ converges to the positive eigenvector $\mathbf{x}_\infty$ corresponding to the unique eigenvalue at $\lambda = 1$. However, this sequence $(\mathbf{x}_k)_{k=0}^\infty$ can experience transient behavior before converging.

In fact, some Markov chains have a certain interesting behavior in which, after an initial period with very little progress towards the steady state, convergence to the steady state occurs quite suddenly. This is known as the *cutoff phenomenon*.

Define the *decay matrix* $A = P - P_\infty$, where P_∞ is the matrix where every column is the steady state vector $\mathbf{x}_\infty$. By induction one can show that $A^k = P^k - P_\infty$ for each $k \geq 1$. Since P^k converges to P_∞ , the powers A^k converge to zero regardless of the norm. But the transient behavior of A is determined by the pseudospectra of A . Careful application of the Kreiss matrix theorem gives a bound for the cutoff phenomenon. For more information on these bounds and how they apply to a variety of problems, including random walks on n -dimensional cubes and card-shuffling, see [JT98].

Application 14.2.11 (Numerical Partial Differential Equations).

An important technique for numerically solving time-dependent partial differential equations is to separate the treatment of space and time. The treatment of the spatial part of the equation can be handled using finite difference, finite element, or spectral methods, which results in a system of coupled ordinary differential equations. We then discretize the system in time, using *time-stepping* formulas. This is called the *method of lines*. The problem with the method of lines is that some choices of discretization of space and time are unstable numerically.

The pseudospectrum plays an important role in determining when a given discretization is stable. In fact there are necessary and sufficient conditions for numerical stability that are expressed in terms of the pseudospectrum. For example, if $L_{\Delta t}$ is the differential operator for the spatial component, and G is a function that characterizes the time-stepping scheme used, and $A_{\Delta t} = G(\Delta t L_{\Delta t})$, then we have

$$\mathbf{v}^{n+1} = G(\Delta t L_{\Delta t}) \mathbf{v}^n,$$

where $\mathbf{v}^n$ is the solution at time n . Determining whether the discretization is stable boils down to showing that $\|A_{\Delta t}^n\| \leq C$ for all $n \in \mathbb{N}$ and Δt with $0 \leq n\Delta t \leq T$ for some fixed function C and all sufficiently small Δt . The Kreiss matrix theorem gives bounds on $\|A_{\Delta t}^n\|$, from which we deduce the stability properties of the discretization. For further details refer to [TE05, RT90, RT92].

14.2.3 Preconditioning

In Section 13.1 we describe three iterative methods for solving linear systems; that is, $A\mathbf{x} = \mathbf{b}$, where $A \in M_n(\mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$ are given. The iterative methods are of the form $\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}$ for some $B \in M_n(\mathbb{F})$ and $\mathbf{c} \in \mathbb{F}^n$. A necessary and sufficient condition for convergence is that the eigenvalues of B be contained in the open unit disk (see Theorem 13.1.1). If for a given $\mathbf{x}_0$ the sequence converges to some $\mathbf{x}_\infty$, the limit satisfies $\mathbf{x}_\infty = B\mathbf{x}_\infty + \mathbf{c}$, which means that the error terms $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}_\infty$ satisfy

$$\mathbf{e}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}_\infty = B(\mathbf{x}_k - \mathbf{x}_\infty),$$

or, equivalently, $\mathbf{e}_{k+1} = B^k(\mathbf{x}_0 - \mathbf{x}_\infty)$. In other words, $\|\mathbf{e}_{k+1}\| \leq \|B^k\| \|\mathbf{x}_0 - \mathbf{x}_\infty\|$. However, if $\|B^k\|$ exhibits transient behavior, the convergence may take a long time to be realized, thus rendering the iterative method of little or no use. Therefore, for this method to be useful, not just the eigenvalues of B must be contained in the unit disk, but also the pseudospectrum must be sufficiently well behaved.

But even if the pseudospectrum of A is not well behaved, not all hope is lost. An important technique is to *precondition* the system $A\mathbf{x} = \mathbf{b}$ by multiplying both sides by a nonsingular matrix M^{-1} so that the resulting linear system $M^{-1}A\mathbf{x} = M^{-1}\mathbf{b}$ has better spectral and transient behavior for its corresponding iterative method.

Of course $M = A$ is the ultimate preconditioner—the resulting linear system is $\mathbf{x} = A^{-1}\mathbf{b}$, which is the solution. However, if we knew A^{-1} , we wouldn't be having this conversation. Also, the amount of work involved in computing A^{-1} is more than the amount of work required to solve the system. So our goal is to find a preconditioner M^{-1} that's close to the inverse of A but is easy to compute and leaves the resulting iterative method with good pseudospectral properties.

If we write $A = M - N$, then $M^{-1}A = I - M^{-1}N$, and the iterative method of Section 13.1 then becomes

$$\mathbf{x}_{k+1} = M^{-1}N\mathbf{x}_k + M^{-1}\mathbf{b}.$$

Hence we need the pseudospectral properties of $B = I - M^{-1}N$ to be well behaved. That is, we want the Kreiss constant of B to be as small as possible. Of course, the overall rate of convergence is determined by the spectral radius, which is bounded by $\|B\|$, so we seek an M that will make $\|B\|$ as small as possible. For example, a preconditioner M that makes $\|B\| = \|I - M^{-1}N\| < 1/2$ would be considered an excellent choice because then the error would satisfy $\|e_{k+1}\| \leq C/2^k$ for some constant C .

14.3 * Proof of the Kreiss Matrix Theorem

14.3.1 Proof of the Kreiss Matrix Theorem

For ease of exposition we break down the proof of the Kreiss theorem into several lemmata, the first of which is due to Spijker. Spijker's lemma is an interesting result that has uses in other areas of mathematics as well, but the proof is rather intricate. We just give the statement of Spijker's lemma here and defer the proof to Section 14.3.2

Lemma 14.3.1 (Spijker's Lemma). *If $\Gamma \subset \mathbb{C}$ is a circle with radius $\rho > 0$ and $r(z) = p(z)/q(z)$ is a rational function of order $n \in \mathbb{N}$, meaning p, q are polynomials over $\mathbb{C}$ of degree at most n , with $q \neq 0$ on Γ , then*

$$\int_{\Gamma} |r'(z)| |dz| \leq 2\pi n \sup_{z \in \Gamma} |r(z)|. \quad (14.13)$$

Lemma 14.3.2. *Given $A \in M_n(\mathbb{C})$, let $r(z) = \mathbf{u}^H R(z) \mathbf{v}$ for some unit vectors $\mathbf{u}$ and $\mathbf{v}$, and let $R(z)$ be the resolvent of A . If $\Gamma = \{z \in \mathbb{C} \mid |z| = 1 + (k+1)^{-1}\}$ for some $k \in \mathbb{N}$, then*

$$\sup_{z \in \Gamma} |r(z)| \leq (k+1)K(A). \quad (14.14)$$

Proof. Using Exercise 4.33, we note that $r(z) \leq \|R(z)\|$. Hence,

$$\begin{aligned} K(A) &= \sup_{|z| > 1} (|z| - 1) \|R(z)\| \geq \sup_{|z| > 1} (|z| - 1) |r(z)| \\ &\geq \sup_{z \in \Gamma} (|z| - 1) |r(z)| = \sup_{z \in \Gamma} (k+1)^{-1} |r(z)|. \quad \square \end{aligned}$$

Now we can state and prove a lemma that will give the right-hand side of (14.12).

Lemma 14.3.3. *If $A \in M_n(\mathbb{C})$, then*

$$\sup_{k \in \mathbb{N}} \|A^k\| \leq enK(A). \quad (14.15)$$

Proof of the Kreiss Matrix Theorem. Let $\Gamma = \{z \in \mathbb{C} \mid |z| = 1 + (k+1)^{-1}\}$ for some $k \in \mathbb{N}$. Let $\mathbf{u}$ and $\mathbf{v}$ be arbitrary unit vectors and define $r(z) = \mathbf{u}^H R(z) \mathbf{v}$ as in Lemma 14.3.2. By the spectral resolution formula (Theorem 12.4.6), we have

$$A^k = \frac{1}{2\pi i} \oint_{\Gamma} z^k R(z) dz,$$

and thus

$$\mathbf{u}^H A^k \mathbf{v} = \mathbf{u}^H \left(\frac{1}{2\pi i} \oint_{\Gamma} z^k R(z) dz \right) \mathbf{v} = \frac{1}{2\pi i} \oint_{\Gamma} z^k \mathbf{u}^H R(A, z) \mathbf{v} dz = \frac{1}{2\pi i} \oint_{\Gamma} z^k r(z) dz.$$

Integrating by parts gives

$$\mathbf{u}^H A^k \mathbf{v} = \frac{-1}{2\pi i(k+1)} \oint_{\Gamma} z^{k+1} r'(z) dz = \frac{i}{2\pi(k+1)} \oint_{\Gamma} z^{k+1} r'(z) dz.$$

On the contour Γ we have

$$|z|^{k+1} = \left(1 + \frac{1}{1+k} \right)^{k+1} < e.$$

Using Spijker's lemma (Lemma 14.3.1) and Lemma 14.3.2, we have

$$\begin{aligned} |\mathbf{u}^H A^k \mathbf{v}| &\leq \frac{1}{2\pi(k+1)} \oint_{\Gamma} |z|^{k+1} |r'(z)| |dz| \\ &\leq \frac{e}{2\pi(k+1)} \oint_{\Gamma} |r'(z)| |dz| \\ &\leq \frac{en}{k+1} \sup_{z \in \Gamma} |r(z)| \\ &\leq enK(A). \end{aligned}$$

Using Exercise 4.33, this gives $\|A^k\| \leq enK(A)$. Since this holds for all $k \in \mathbb{N}$, we have (14.15). $\square$

14.3.2 Proof of Spijker's Lemma

Spijker's lemma is a consequence of the following five lemmata. The fourth lemma is hard, but the others are relatively straightforward.

Lemma 14.3.4. *Let $p(z)$ be a polynomial of degree $n \in \mathbb{N}$. The restriction of $z^n \overline{p(z)}$ to the circle $S = \{z \in \mathbb{C} : |z| = \rho\}$, where $\rho > 0$, is equivalent to the restriction to S of some polynomial of degree at most n .*

Proof. Let $p(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + a_0$. Assume $z(t) = \rho e^{it}$. Note that $\overline{z(t)} = \rho e^{-it} = \rho^2 z(t)^{-1}$. Thus,

$$\overline{p(z)} = a_n \rho^{2n} z^{-n} + a_{n-1} \rho^{2(n-1)} z^{-(n-1)} + \dots + a_1 \rho^2 z^{-1} + a_0.$$

It follows that

$$z^n \overline{p(z)} = a_n \rho^{2n} + a_{n-1} \rho^{2(n-1)} z + \dots + a_1 \rho^2 z^{n-1} + a_0 z^n. \quad \square$$

Lemma 14.3.5. Assume $f(t) = r(\rho e^{it})$. We have that

$$\int_S |r'(t)| |dz| = \int_0^{2\pi} |f'(t)| dt. \quad (14.16)$$

Proof. Since $f'(t) = \rho i e^{it} r'(\rho e^{it})$, we have $|f'(t)| = \rho |r'(\rho e^{it})|$, and since $z = \rho e^{it}$ we have that $|dz| = \rho dt$. Thus,

$$\int_S |r'(t)| |dz| = \int_0^{2\pi} \frac{|f'(t)|}{\rho} \rho dt = \int_0^{2\pi} |f'(t)| dt. \quad \square$$

Lemma 14.3.6. Let $g'(t) = |f'(t)| \cos \omega(t)$ and $h'(t) = |f'(t)| \sin \omega(t)$ be the real and imaginary parts of $|f'(t)|$, respectively. The following equality holds:

$$|f'(t)| = \frac{1}{4} \int_0^{2\pi} |g'(t) \cos \theta + h'(t) \sin \theta| d\theta. \quad (14.17)$$

Proof. We have that

$$\begin{aligned} \int_0^{2\pi} |g'(t) \cos \theta + h'(t) \sin \theta| d\theta &= \int_0^{2\pi} |f'(t)| |\cos \omega(t) \cos \theta + \sin \omega(t) \sin \theta| d\theta \\ &= |f'(t)| \int_0^{2\pi} |\cos(\omega(t) - \theta)| d\theta \\ &= 4|f'(t)|. \quad \square \end{aligned}$$

Lemma 14.3.7. For fixed $\theta \in [0, 2\pi]$, define $F_\theta(t) = g(t) \cos \theta + h(t) \sin \theta$. We have

$$\int_0^{2\pi} |F'_\theta(t)| dt \leq 4n \sup_{t \in [0, 2\pi]} |F_\theta(t)|. \quad (14.18)$$

Proof. Since $F'_\theta(t) = g'(t) \cos \theta + h'(t) \sin \theta$ is derived from a nontrivial rational function, it has finitely many distinct roots $0 \leq t_0 < t_1 < \dots < t_{k-1} < 2\pi$. Write $t_k = 2\pi$. Thus, $|F'_\theta(t)| > 0$ on each interval (t_{j-1}, t_j) for $j = 1, 2, \dots, k$. If $F'_\theta(t) < 0$ on (t_{j-1}, t_j) , then

$$\int_{t_{i-1}}^{t_i} |F'_\theta(t)| dt = - \int_{t_{i-1}}^{t_i} F'_\theta(t) dt = -(F_\theta(t_i) - F_\theta(t_{i-1})) = |F_\theta(t_i) - F_\theta(t_{i-1})|.$$

If instead $F'_\theta(t) > 0$ on (t_{j-1}, t_j) , then

$$\int_{t_{i-1}}^{t_i} |F'_\theta(t)| dt = \int_{t_{i-1}}^{t_i} F'_\theta(t) dt = (F_\theta(t_i) - F_\theta(t_{i-1})) = |F_\theta(t_i) - F_\theta(t_{i-1})|.$$

Thus,

$$\int_0^{2\pi} |F'_\theta(t)| dt = \sum_{i=1}^k \int_{t_{i-1}}^{t_i} F'_\theta(t) dt = \sum_{i=1}^k |F_\theta(t_i) - F_\theta(t_{i-1})| \leq 2k \sup_{t \in [0, 2\pi]} |F_\theta(t)|.$$

It suffices to show that $k \leq 2n$. Note that

$$\begin{aligned} 2F_\theta(t) &= g(t)(e^{i\theta} + e^{-i\theta}) - ih(t)(e^{i\theta} - e^{-i\theta}) \\ &= e^{i\theta}(g(t) - ih(t)) + e^{-i\theta}(g(t) + ih(t)) \\ &= e^{i\theta}\overline{r(e^{it})} + e^{-i\theta}r(e^{it}). \end{aligned}$$

Recall that $z(t) = \rho e^{it}$. Multiplying both sides by $z^n q(z) \overline{q(z)}$, we have

$$2z^n F_\theta(t) q(z) \overline{q(z)} = e^{i\theta} q(z) (z^n \overline{p(z)}) + e^{-i\theta} p(z) (z^n \overline{q(z)}).$$

Since $q(z)$ is nonzero on S and, by Lemma 14.3.4, the right-hand side is a polynomial on S of degree at most $2n$, we can conclude that $F_\theta(t)$ has at most $2n$ roots. Therefore $k \leq 2n$. $\square$

Lemma 14.3.8. *For fixed $\theta \in [0, 2\pi]$ we have*

$$\sup_{t \in [0, 2\pi]} |F_\theta(t)| dt \leq \sup_{z \in S} |r(z)|. \quad (14.19)$$

Proof. Using the Cauchy–Schwarz inequality (Proposition 3.1.17), we have

$$\begin{aligned} |F_\theta(t)| &= |g(t) \cos \theta + h(t) \sin \theta| \\ &= \left\| \begin{bmatrix} g(t) \\ h(t) \end{bmatrix} \cdot \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \right\| \\ &\leq \sqrt{g(t)^2 + h(t)^2} \sqrt{\cos^2 \theta + \sin^2 \theta} \\ &= \sqrt{g(t)^2 + h(t)^2} = |r(t)|. \quad \square \end{aligned}$$

Now we have all the pieces needed for an easy proof of Spijker's lemma.

Proof of Spijker's Lemma. Using Lemmata 14.3.5–14.3.8, we have

$$\begin{aligned} \int_S |r'(t)| |dz| &= \int_0^{2\pi} |f'(t)| dt \\ &= \int_0^{2\pi} \left(\frac{1}{4} \int_0^{2\pi} |g'(t) \cos \theta + h'(t) \sin \theta| d\theta \right) dt \\ &= \frac{1}{4} \int_0^{2\pi} \left(\int_0^{2\pi} |F'_\theta(t)| dt \right) d\theta \\ &= n \int_0^{2\pi} \left(\sup_{t \in [0, 2\pi]} |F_\theta(t)| \right) d\theta \\ &\leq 2\pi n \sup_{z \in S} |r(z)|. \quad \square \end{aligned}$$

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

14.1. Given $A \in M_n(\mathbb{F})$, show that $\sigma_{\varepsilon_1}(A) \subset \sigma_{\varepsilon_2}(A)$ whenever $0 < \varepsilon_1 < \varepsilon_2$.

14.2. Prove the following: Given $A \in M_n(\mathbb{C})$, there exists $\gamma > 0$ and $M \geq 1$ such that

$$\|A^k\| \leq M\gamma^k \quad \text{for all } k \in \mathbb{N}.$$

Hint: Given $\varepsilon > 0$, choose $\tilde{A} = (r(A) + \varepsilon)^{-1}A$ and use Exercise 12.15.

14.3. Given $A \in M_n(\mathbb{F})$, show that for any $z \in \mathbb{C}$ and $\varepsilon > 0$, we have

$$\sigma_\varepsilon(zI - A) = z - \sigma_\varepsilon(A).$$

14.4. Prove Proposition 14.1.10; that is, show the equivalence (14.4) and (14.2).

14.5. For any $A \in M_n(\mathbb{C})$, any nonzero $c \in \mathbb{C}$, and $\varepsilon > 0$, prove the following:

(i) $\sigma_\varepsilon(A^H) = \overline{\sigma_\varepsilon(A)}$.

(ii) $\sigma_\varepsilon(A + c) = \sigma_\varepsilon(A) + c$.

(iii) $\sigma_{|c|\varepsilon}(A + c) = c\sigma_\varepsilon(A)$.

14.6. Prove: If A is normal, then $r_\varepsilon(A) = r(A) + \varepsilon$.

14.7. Prove: If $A \in M_n(\mathbb{C})$ is normal, then $\|A^k\| = r(A)^k$.

14.8. Given $\varepsilon > 0$, let

$$A = \begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix}.$$

(i) Use induction to prove that

$$A^k = \begin{bmatrix} 1 & \varepsilon k \\ 0 & 1 \end{bmatrix}.$$

(ii) Use Exercise 3.28 to show that $\|A^k\|_2 \rightarrow \infty$ as $k \rightarrow \infty$. Hint: It is trivial to compute the 1-norm and ∞ -norm of A^k .

- 14.9. Given $A \in M_n(\mathbb{C})$, assume that $r(A) = 1$, but that the eigennilpotent D_λ is zero for all λ such that $|\lambda| = 1$. Prove one direction of Theorem 14.2.1(iii) by completing following steps:

(i) Use the spectral decomposition to write

$$A = \sum_{|\lambda|=1} \lambda P_\lambda + B,$$

where $r(B) < 1$ and $P_\lambda B = B P_\lambda = 0$ for all $\lambda \in \sigma(A)$ satisfying $|\lambda| = 1$.

(ii) Show that $A^k = \sum_{|\lambda|=1} \lambda^k P_\lambda + B^k$.

(iii) Use the triangle inequality to show that

$$\|A^k\| \leq \sum_{|\lambda|=1} \|P_\lambda\| + \|B^k\|.$$

Then take the limit as $k \rightarrow \infty$.

- 14.10. Given $A \in M_n(\mathbb{C})$, assume there exists an eigenvalue of modulus one that has an eigennilpotent. Prove that $\|A^k\| \rightarrow \infty$ as $k \rightarrow \infty$. This is the other direction of Theorem 14.2.1(iii).

- 14.11. Show that $(1 + \frac{1}{1+k})^{k+1}$ is strictly monotonic in $k \in \mathbb{N}$ by completing the following steps:

(i) Let $f(x) = x^{k+1}$ and use the mean value theorem for $0 \leq a < b$ to get

$$b^{k+1} - a^{k+1} < (k+1)b^k(b-a).$$

(ii) By expanding the previous expression, show that

$$a^{k+1} > b^k((k+1)a - kb).$$

(iii) Complete the proof by choosing $a = 1 + (k+1)^{-1}$ and $b = 1 + k^{-1}$.

- 14.12. Given $A \in M_n(\mathbb{F})$, prove that the Kreiss constant $K(A)$ is the smallest $C > 0$ such that

$$\|(zI - A)^{-1}\| \leq \frac{C}{|z| - 1} \quad \text{for all } |z| > 1.$$

- 14.13. Given $A \in M_n(\mathbb{F})$ and the resolvent estimate in the previous problem, show that for a contour of radius $1 + \varepsilon$, centered at the origin with $\varepsilon > 0$, we have that

$$\|A^k\| \leq \frac{(1 + \varepsilon)^k}{\varepsilon} K(A).$$

Then, for each k , choose ε to minimize the right-hand side, leaving the bound

$$\|A^k\| \leq k \left(1 + \frac{1}{k}\right)^k K(A) \leq keK(A).$$

Since this grows without bound, this isn't a very useful bound. It's for this reason that the integration-by-parts step in Lemma 14.3.3 is so critical.

14.14. By integrating by parts twice in Lemma 14.3.3, show that

$$\|A^k\| \leq \frac{2en^2}{(k+1)(k+2)} \sup_{z \in S} |r'(z)|.$$

Hint: The derivative of a rational function is a rational function.

14.15. If it is known a priori that the numerator and denominator of the rational function $r(z)$ have degrees $n-1$ and n , respectively, can Spijker's lemma be improved, and if so, what does this mean for the Kreiss matrix theorem?

Notes

Much of our treatment of pseudospectra was inspired by [Tre92, Tre97, TE05], and Exercise 14.5 is from [TE05]. All figures showing contour plots of pseudospectra were created using Eigtool [Wri02].

For more about the spectral condition number and the choice of V in (14.6) and in the Bauer–Fike theorem (Corollary 14.1.15), see [Dem83, Dem82, JL97].

Finally, for a fascinating description of the history and applications of Spijker's lemma, we strongly recommend the article [WT94].

15 Rings and Polynomials

Հանրահայտը. հանրահայտը
հանրահայտը. սթեյնշտեյն

—Sauron

Ring theory has many applications in computing, counting, cryptography, and communications. Rings are sets with two operations, “+” and “·” (usually called *addition* and *multiplication*, respectively). In many ways rings are like vector spaces, and much of our treatment of the theory of rings will mirror our treatment of vector spaces at the beginning of this book. The prototypical ring is the set $\mathbb{Z}$ with addition and multiplication as the two operations. But there are many other interesting and useful examples of rings, including $\mathbb{F}[x]$, $M_n(\mathbb{F})$, $C(U; \mathbb{F})$, and $\mathcal{B}(X)$ for any normed linear space X . Ring theory shows that many results that are true for the integers also apply to these rings.

We begin with a survey of the basic properties of rings, with a special focus on the similarities between rings and vector spaces. We then narrow our focus to a special class of rings called *Euclidean domains*, and to quotients of Euclidean domains. These include rings of polynomials in one variable and the ring of all the matrices that can be formed by applying polynomials to one given matrix. A Euclidean domain is a ring in which the Euclidean algorithm holds, and we show in the third section that this also implies the fundamental theorem of arithmetic (unique prime factorization) holds in these rings. Applying this to the ring of polynomials, we get an alternative form of the fundamental theorem of algebra.

In the fourth section we treat homomorphisms, which are the ring-theoretic analogue of linear transformations. We show there are strong parallels between maps of rings (homomorphisms) and maps of vector spaces (linear transformations). Again the kernel plays a fundamental role, and the first and second isomorphism theorems for rings both hold.⁵¹

⁵¹There is also a third isomorphism theorem both for rings and for vector spaces. We do not cover that theorem here, but you can find it in almost any standard text on ring theory.

One of the most important results of this chapter is the Chinese remainder theorem (CRT), which we prove in Section 15.6. It is a powerful tool with many applications in both pure and applied mathematics. In the remainder of the chapter we focus primarily on the important implications of the CRT for partial fraction decomposition, polynomial interpolation, and spectral decomposition of operators. We conclude the chapter by describing a remarkable connection between Lagrange interpolants and the spectral decomposition of a matrix (see Section 15.7.3).

15.1 Definition and Examples

15.1.1 Definition and Basic Properties

Definition 15.1.1. A ring is a set R of elements with two operations: addition, mapping $R \times R$ to R , and denoted by $(x, y) \mapsto x + y$; and multiplication, mapping $R \times R$ to R , and denoted by $(x, y) \mapsto x \cdot y$ or just $(x, y) \mapsto xy$. These operations must satisfy the following properties for all $x, y, z \in R$:

- (i) *Commutativity of addition:* $x + y = y + x$.
- (ii) *Associativity of addition:* $(x + y) + z = x + (y + z)$.
- (iii) *Existence of an additive identity:* There exists an element $0 \in R$ such that $0 + x = x$.
- (iv) *Existence of an additive inverse:* For each $x \in R$ there exists an element, denoted $-x$, in R such that $x + (-x) = 0$.
- (v) *First distributive law:* $z(x + y) = zx + zy$.
- (vi) *Second distributive law:* $(x + y)z = xz + yz$.
- (vii) *Associativity of multiplication:* $(xy)z = x(yz)$.

Remark 15.1.2. A ring is like a vector space in many ways. The main differences are in the nature of the multiplication. In a vector space we multiply by scalars that come from a field—outside the vector space. But in a ring we multiply by elements inside the ring.

Example 15.1.3. Even if you have seen rings before, you should familiarize yourself with the following examples, since most of them are very common in mathematics and will arise repeatedly in many different contexts.

- (i) The integers $\mathbb{Z}$ form a ring.
- (ii) The rationals $\mathbb{Q}$, the reals $\mathbb{R}$, and the complex numbers $\mathbb{C}$ each form a ring (with the usual operations of addition and multiplication).

- (iii) Fix a positive integer n . The set $\mathbb{Z}_n = \{[0], \dots, [n-1]\}$ of equivalence classes mod n , as described in Example A.1.18(i), forms a ring with operations $\boxplus$ and $\boxdot$, as described in Examples A.2.16(i) and A.2.16(ii).
- (iv) The set $C^\infty((a, b); \mathbb{F})$ of smooth functions from (a, b) to $\mathbb{F}$ forms a ring, where $+$ is pointwise addition $(f+g)(x) = f(x) + g(x)$ and $\cdot$ is pointwise multiplication $(f \cdot g)(x) = f(x)g(x)$.
- (v) Given a vector space X , the set $\mathcal{B}(X)$ of bounded linear operators on X is a ring, where $+$ is the usual addition of operators and $\cdot$ is composition of operators. In particular, the set of $n \times n$ matrices $M_n(\mathbb{F})$ is a ring, with the usual matrix addition and matrix multiplication.
- (vi) Given a set S , the power set $\mathcal{P}(S)$ forms a ring, where $+$ is the symmetric difference and $\cdot$ is intersection:

$$A + B = (A \cup B) \setminus (A \cap B) \quad \text{and} \quad A \cdot B = A \cap B.$$

The additive identity is $\emptyset$, and any subset A is its own additive inverse $-A = A$.

Warning: Some people use $+$ in this setting to denote union $\cup$. This definition of $+$ does not satisfy the axioms of a ring.

- (vii) The set $\{\text{True}, \text{False}\}$ of Boolean truth values forms a ring, where $+$ is the operation of *exclusive OR* (XOR); that is, $a + b = \text{True}$ if and only if *exactly one* of a and b is True. Multiplication $\cdot$ is the operation AND. The additive identity 0 is the element False, and the additive inverse of any element is itself: $-a = a$.

Again, you should be aware that many people use $+$ to denote inclusive OR, but inclusive OR will not satisfy the axioms of a ring.

Remark 15.1.4. Note that the definition of a ring requires addition to be commutative but does not require (or forbid) that multiplication be commutative. Perhaps the most common example of a ring with noncommutative multiplication is the ring $M_n(\mathbb{F})$.

Definition 15.1.5. A ring R is commutative if $ab = ba$ for all $a, b \in R$.

Example 15.1.6.

- (i) Given any commutative ring R , the set

$$R[x] = \{a_0 + a_1x + \dots + a_kx^k \mid k \in \mathbb{N}, \text{ all } a_i \in R\}$$

of all polynomials in the variable x with coefficients in R forms a commutative ring, where the operations $+$ and $\cdot$ are the usual addition and multiplication of polynomials.

Since $R[x]$ is itself a ring, we may repeat the process with a new variable y to get that $R[x, y] = R[x][y]$ is also a ring, as is $R[x, y, z]$, and so forth.

- (ii) Given any commutative ring R , the set

$$R[[x]] = \left\{ \sum_{i=0}^{\infty} a_i x^i \mid a_i \in R \text{ for all } i \in \{0, 1, \dots\} \right\}$$

of *formal power series* in the indeterminate x is a commutative ring. Elements of $R[[x]]$ need not converge. This is why they are called *formal* power series.

- (iii) The set

$$R[x, x^{-1}] = \left\{ \sum_{i=N}^M a_i x^i \mid N \leq M \in \mathbb{Z}, a_i \in R \text{ for all } i \in \{N, \dots, M\} \right\}$$

of *Laurent polynomials* in the indeterminate x with coefficients in R is a commutative ring if R is a commutative ring. Notice here that N may be negative, but N and M must be finite.

- (iv)  If $A \in M_n(\mathbb{F})$ is a square matrix, the set

$$\mathbb{F}[A] = \left\{ \sum_{i=0}^N c_i A^i \mid N \in \mathbb{N}, \text{ all } c_i \in \mathbb{F} \right\} \subset M_n(\mathbb{F})$$

of polynomials in A with coefficients in $\mathbb{F}$ is a ring.

Note that while the symbol x in the rings $\mathbb{F}[x]$ and $\mathbb{F}[[x]]$ and $\mathbb{F}[x, x^{-1}]$ is a *formal symbol*, the A in $\mathbb{F}[A]$ is a specific matrix. So the elements of $\mathbb{F}[x]$, for example, are all the formal expressions of the form $a_0 + a_1x + \dots + a_nx^n$, but the elements of $\mathbb{F}[A]$ are a specific set of matrices. For example, if $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, then $A^n = 0$ for $n > 1$, and so $\mathbb{C}[A] = \{c_1I + c_2A \mid c_1, c_2 \in \mathbb{C}\} = \left\{ \begin{bmatrix} c_1 & c_2 \\ 0 & c_1 \end{bmatrix} \mid c_1, c_2 \in \mathbb{C} \right\}$ is the set of all upper-triangular matrices with constant diagonal.

The ring $\mathbb{F}[A] \subset M_n(\mathbb{F})$ is a *commutative* ring because any power of A commutes with any other power of A , despite the fact that it is a subring of $M_n(\mathbb{F})$, which is not commutative.

Nota Bene 15.1.7. As with vector spaces, a subtle point that is sometimes missed, because it is not a separate item in the numbered list of axioms, is that the definition of an operation requires that the operations of addition and multiplication take their values in R . That is, R must be *closed under addition and multiplication*.

Unexample 15.1.8.

- (i) The natural numbers $\mathbb{N}$ with the usual operations of $+$ and $\cdot$ do not form a ring because every nonzero element fails to have an additive inverse.
- (ii) The odd integers $\mathbb{O}$ with the usual operations $+$ and $\cdot$ do not form a ring because the operation of $+$ takes two odd integers and returns an even. That is, the operation $+$ is *not an operation on odd integers*, since an operation on $\mathbb{O}$ is a function $\mathbb{O} \times \mathbb{O} \rightarrow \mathbb{O}$, but the range of $+$ is not $\mathbb{O}$. Instead of saying that $+$ is not an operation on $\mathbb{O}$, many people say *the set $\mathbb{O}$ is not closed under addition*.

Proposition 15.1.9. *Let R be a ring. If $x, y \in R$, then the following hold:*

- (i) *The additive identity 0 is unique, that is, $x + y = x$ implies $y = 0$.*
- (ii) *Additive inverses are unique, that is, $x + y = 0$ implies $y = (-x)$.*
- (iii) $0 \cdot x = 0$.
- (iv) $(-y) \cdot x = -(y \cdot x) = y \cdot (-x)$.

Proof. The proof is identical to the case of vector spaces (Proposition 1.1.7), except for (iv), and even in that case it is similar. To see (iv) note that for each $x, y \in R$, we have $0 = 0 \cdot x = (y + (-y)) \cdot x = y \cdot x + (-y) \cdot x$. Hence, (ii) implies that $(-y) \cdot x = -(y \cdot x)$. A similar argument shows that $-(y \cdot x) = y \cdot (-x)$. $\square$

Remark 15.1.10. Since the expression $x + (-y)$ determines a well-defined (unique) element, we usually use the shorthand $x - y$ to denote $x + (-y)$. We also usually write xy instead of $x \cdot y$.

Although the axioms of a ring require an additive identity element 0 , they do not require the existence of a multiplicative identity (usually denoted 1).

Definition 15.1.11. *Let u in a ring R have the property that for every $r \in R$ we have $ur = ru = r$. In this case u is called unity and is usually denoted 1 .*

Proposition 15.1.12. *If R has a unity element, then it is unique.*

Proof. Suppose that R contains two elements u and u' such that $ur = ru = r$ and $u'r = ru' = r$ for every $r \in R$. We have $u = uu'$ by the unity property of u' , but also $uu' = u'$ by the unity property of u . Hence $u = u'$. $\square$

Multiplicative inverses are also not required in a ring, but when they exist, they are very useful.

Definition 15.1.13. For any ring R with unity 1, an element $a \in R$ is invertible⁵² if there exists $b \in R$ such that $ab = 1 = ba$. We usually denote this b by a^{-1} .

Proposition 15.1.14. In any ring R with unity 1, if an element $a \in R$ has a multiplicative inverse, then its inverse is unique.

Proof. If there exist elements $b, b' \in R$ such that $ab = 1$ and $b'a = 1$, then we have

$$b = 1b = (b'a)b = b'(ab) = b'. \quad \square$$

Example 15.1.15. In the ring $\mathbb{Z}$, the only invertible elements are 1 and -1 , but in the ring $\mathbb{Q}$ every nonzero element is invertible.

15.1.2 Ideals

Roughly speaking, an ideal is to a ring what a vector subspace is to a vector space. But this analogy is incomplete because an ideal is not just a subring (a subset that is also a ring)—it must satisfy some additional conditions that are important for building quotients of rings.

Definition 15.1.16. Let R be a ring. A nonempty subset $I \subset R$ is an ideal of R if the operations of addition and multiplication in R satisfy the following properties:

- (i) For any $x, y \in I$, we have $x + y \in I$ and $x - y \in I$.
- (ii) For any $r \in R$ and any $x \in I$, we have $rx \in I$ and $xr \in I$.

Remark 15.1.17. Of course the first condition is just saying that I is closed under addition and subtraction, but the second condition is new to us— I must be *closed under multiplication by any ring element*. Roughly speaking, this is analogous to the condition that vector subspaces should be closed under scalar multiplication, but here the analogues of scalars are elements of the ring R .

Example 15.1.18. The set $\mathbb{E} = 2\mathbb{Z}$ of even integers is an ideal of the integers $\mathbb{Z}$, because it is closed under addition, subtraction, and multiplication by any integer.

⁵²It is common to call an invertible element of a ring a *unit*, but this term is easily confused with *unity*, so we prefer not to use it.

Unexample 15.1.19. The set $\mathbb{E} = 2\mathbb{Z}$ is not an ideal of the rationals $\mathbb{Q}$, because it is not closed under multiplication by an element from $\mathbb{Q}$. For example, $2 \in \mathbb{E}$, but $\frac{1}{3} \cdot 2 = \frac{2}{3} \notin \mathbb{E}$.

This unexample, combined with Example 15.1.18, shows that talking about an ideal makes no sense without specifying what ring it is an ideal of.

Proposition 15.1.20. *Let R be a ring. Any ideal $I \subset R$ of a ring R is itself a ring, using the same operations $+$ and $\cdot$.*

Proof. From the definition of ideal, we have that I is closed under addition and multiplication, and hence those operations are operations on I . The properties of associativity for $+$ and $\cdot$, as well as commutativity of $+$ and distributivity all follow immediately from the fact that they hold in the larger ring R .

All that remains is to check that the additive identity 0 is in I and that every element of I has an additive inverse in I . But these both follow by closure under subtraction. First, for any $x \in I$ we have $0 = x - x \in I$. Now since $0 \in I$, we have $-x = 0 - x \in I$. $\square$

Remark 15.1.21. Example 15.1.19 shows that the converse to the previous proposition is false: not every subring is an ideal.

Proposition 15.1.22. *To check that a nonempty subset I of a ring R is an ideal, it suffices to check that*

- (i) $rx \in I$ and $xr \in I$ for all $r \in R$ and $x \in I$ (closure under multiplication by any ring element), and
- (ii) $x - y \in I$ for all $x, y \in I$ (closure under subtraction).

Proof. If I is closed under subtraction, then given any $x, y \in I$, we have $y - y = 0 \in I$, and thus $-y = 0 - y \in I$. So we have $x + y = x - (-y) \in I$. $\square$

The next proposition is immediate, and in fact it may even seem more difficult than checking the definition directly, but it makes many proofs much cleaner; it also makes the similarity to vector subspaces clearer.

Proposition 15.1.23. *If R is a ring with unity, then to check that a nonempty subset $I \subset R$ is an ideal, it suffices to check that $ax + by \in I$ and $xa + yb \in I$ for every $a, b \in R$ and every $x, y \in I$. If R is commutative, it suffices to check just that $ax + by \in I$ for every $a, b \in R$ and every $x, y \in I$.*

Example 15.1.24.

- (i) Every ring R is an ideal of itself, but it may not be an ideal of a larger ring.

- (ii) The set $\{0\}$ is an ideal in any ring.
- (iii) For any $p \in \mathbb{R}$, the set $\mathfrak{m}_p = \{f \in C^\infty(\mathbb{R}; \mathbb{R}) \mid f(p) = 0\}$ of all functions that vanish at p is an ideal in the ring $C^\infty(\mathbb{R}; \mathbb{R})$.
- (iv) If $n, d, k \in \mathbb{Z}$ are positive integers with $dk = n$, then the set of $(\text{mod } n)$ equivalence classes $d\mathbb{Z}_n = \{[0], [d], [2d], \dots, [(k-1)d]\}$ is an ideal of $\mathbb{Z}_n$.
- (v) For any ring R , the set

$$I = \{a_1x + \dots + a_kx^k \mid k \in \mathbb{N}, a_i \in R \text{ for all } i \in \{1, \dots, k\}\}$$

of all polynomials with zero constant term is an ideal in the polynomial ring $R[x]$.

- (vi) The matrices $M_n(2\mathbb{Z}) \subset M_n(\mathbb{Z})$ of matrices with even integer entries forms an ideal in the ring of matrices with integer entries.

Unexample 15.1.25.

- (i) The set $\mathbb{O}$ of odd integers is not an ideal in $\mathbb{Z}$ because it is not closed under subtraction or addition.
- (ii) The only ideals of $\mathbb{Q}$ are $\{0\}$ and $\mathbb{Q}$ itself. Any ideal $I \subset \mathbb{Q}$ that is not $\{0\}$ must contain a nonzero element $x \in I$. But since $x \neq 0$, we also have $1/x \in \mathbb{Q}$. Let s be any element of $\mathbb{Q}$, and take $r = s/x \in \mathbb{Q}$. Closure under multiplication by elements of $\mathbb{Q}$ implies that $s = (s/x)x = rx \in I$.

15.1.3 Generating Sets and Products of Ideals

Rings are like vector spaces and ideals like subspaces in many ways. In the theory of rings and ideals, generating sets play the same role that spanning sets play for vector subspaces. But one important way that rings differ from vector spaces is that most rings do not have a meaningful analogue of linear independence. So while we can talk about generating sets (like spanning sets), we do not have all the nice properties of a basis, and we cannot define dimension in the same way we do for vector spaces. Nevertheless, many results on spanning sets and quotients of vectors spaces carry over to generating sets and ring quotients. For many of these results, after making the obvious adjustments from vector spaces to rings, the proof is essentially identical to the vector space proof.

Throughout this section, assume that R is a commutative ring and that S is a nonempty subset of R . Most of what we discuss here has a straightforward generalization to noncommutative rings, but the commutative case is simpler and is all we are interested in for now.

Definition 15.1.26. Let R be a commutative ring. The ideal of R generated by S , denoted (S) , is the set of all finite sums of the form

$$a_1x_1 + a_2x_2 + \cdots + a_mx_m, \quad x_i \in S, a_i \in R, m \in \mathbb{Z}^+. \quad (15.1)$$

We call such a sum an R -linear combination of elements of S . If $(S) = I$ for some ideal I of R , then we say that S is a generating set for I , or equivalently that S generates I .

Proposition 15.1.27. Let R be a commutative ring. For any subset $S \subset R$, the set (S) is an ideal of R .

Proof. If $x, y \in (S)$, then there exists a finite subset $\{x_1, \dots, x_m\} \subset S$, such that $x = \sum_{i=1}^m c_i x_i$ and $y = \sum_{i=1}^m d_i x_i$ for some coefficients $c_1, \dots, c_m$ and $d_1, \dots, d_m$ (some possibly zero). Since $ax + by = \sum_{i=1}^m (ac_i + bd_i)x_i$ is an R -linear combination of elements of S , it follows that $ax + by$ is contained in (S) . Since R is commutative, we also have $xa + yb = ax + by \in (S)$; hence, (S) is an ideal of R . $\square$

Corollary 15.1.28. Let R be a commutative ring. Let $\{I_j\}_{j=1}^n$ be a finite set of ideals in R . The sum $I = I_1 + \cdots + I_n = \{\sum_{j=1}^n a_j \mid a_j \in I_j\} = (I_1 \cup I_2 \cup \cdots \cup I_n)$ is an ideal of R .

Proposition 15.1.29. Let R be a commutative ring. If I is an ideal of R , then $(I) = I$.

Proof. The proof is by induction on the number of terms in an R -linear combination. It is essentially identical to that given for vector spaces (Proposition 1.2.2). The only changes we need to make to that proof are replacing scalars in $\mathbb{F}$ with elements of R . $\square$

Example 15.1.30.

- (i) The ideal (2) in the ring $\mathbb{Z}$ is the set $(2) = \{\dots, -4, -2, 0, 2, 4, \dots\}$ of even integers. More generally, for any $d \in \mathbb{Z}$ the ideal (d) is the set $(d) = \{\dots, -2d, -d, 0, d, 2d, \dots\}$ of all multiples of d .
- (ii) For any ring R , and $x \in R$, the ideal (x) is the set of all multiples of x .
- (iii) The ideal $(6, 9)$ in $\mathbb{Z}$ is the set of all $\mathbb{Z}$ -linear combinations of 6 and 9. Since $3 = 9 - 6$, we have $3 \in (6, 9)$, which implies that $(3) \subset (6, 9)$. But since 6 and 9 are both elements of (3) , we also have $(6, 9) \subset (3)$, and hence $(6, 9) = (3)$.
- (iv) In the polynomial ring $\mathbb{C}[x, y]$, the ideal (x, y) is the set of all polynomials whose constant term is zero.

(v) In the polynomial ring $\mathbb{C}[x, y]$ the ideal $I = (x^2 - y + 1, 2y - 2)$ contains the element $x^2 - y + 1 + \frac{1}{2}(2y - 2) = x^2$, and also the element $\frac{1}{2}(2y - 2) = y - 1$, so $(x^2, y - 1) \subset I$, but also $x^2 - y + 1 = x^2 - (y - 1)$ and $2y - 2 = 2(y - 1)$, so we have $I = (x^2 - 1, y - 1)$.

Nota Bene 15.1.31. A good way to show an ideal is contained in another is to show its generators lie in the other ideal. To show that two ideals are equal, show that the generators of each are contained in the other ideal.

Example 15.1.32. In a commutative ring R with unity 1, for any invertible element $u \in R$, the ideal (u) , generated by u , is the entire ring; that is, $(u) = R$. To see this, note that for any element $r \in R$, we have $r = (ru^{-1})u \in (u)$.

Nota Bene 15.1.33. When talking about elements of a ring, parentheses are still sometimes used to specify order of operations, so you will often encounter confusing notation like $(x^2 - (y - 1))$. If $(x^2 - (y - 1))$ is supposed to mean an element of R , then $(x^2 - (y - 1)) = x^2 - y + 1$, but if it is supposed to be an ideal, then $(x^2 - (y - 1)) = (x^2 - y + 1)$ means the ideal generated by $x^2 - y + 1$, that is, the set of all elements of R that can be written as a multiple of $x^2 - y + 1$. The meaning should be clear from the context, but it takes some getting used to; so you'll need to pay careful attention to identify the intended meaning. Although this notation is confusing and is certainly suboptimal, it is the traditional notation, so you should become comfortable with it.

The proofs of the remaining results in this section are all essentially identical to their vector space counterparts in Section 1.5. The details are left as exercises.

Proposition 15.1.34. *The intersection of a collection $\{I_\alpha\}_{\alpha \in J}$ of ideals of R is an ideal of R .*

Proposition 15.1.35. *Supersets of generating sets are also generating sets, that is, if R is a commutative ring, if $(S) = I$, and if $S \subset S' \subset I$, then $(S') = I$.*

Theorem 15.1.36. *Let R be a commutative ring. The ideal generated by S is the smallest ideal of R that contains S , or, in other words, it is the intersection of all ideals of R that contain S .*

Proposition 15.1.37. *If $I_1, \dots, I_n$ are ideals in a commutative ring R , then the product of these ideals, defined by $\prod_{i=1}^n I_i = \{\sum_{j=1}^k (\prod_{i=1}^n a_{ji}) \mid a_{ji} \in I_i\}$, is an ideal of R and a subset of $\bigcap_{i=1}^n I_i$.*

15.2 Euclidean Domains

Some of the most important rings in applications are the rings⁵³ $\mathbb{Z}$ and $\mathbb{F}[x]$. Some key properties of these rings are that they are all commutative and satisfy a form of the division property; thus they have a form of the Euclidean algorithm. These properties allow us to talk meaningfully about primes and divisibility.

Rings that have these properties are called *Euclidean domains*. Later we will use the properties of Euclidean domains to develop the partial fraction decomposition for rational functions and to better understand polynomial interpolation.

15.2.1 Euclidean Domains and Polynomial Rings

Euclidean domains are commutative rings where the division algorithm and the Euclidean algorithm work. In this section we give a careful definition of Euclidean domains and then prove that polynomials in one variable form a Euclidean domain.

The key to the Euclidean algorithm for the integers is the fact that size (absolute value) of the remainder r is smaller than that of the divisor b . In a general ring, the absolute value may not make sense, but in many cases we can find a function that will take the place of the absolute value. We call such a function a *valuation*.

Definition 15.2.1. *A ring R is a Euclidean domain if the following hold:*

- (i) R has a multiplicative identity element 1.
- (ii) R is commutative; that is, for every $a, b \in R$ we have $ab = ba$.
- (iii) R has no zero divisors; that is, for any $a, b \in R$, if $ab = 0$, then $a = 0$ or $b = 0$.
- (iv) There exists a function $v : R \setminus \{0\} \rightarrow \mathbb{N}$ (called a valuation) such that
 - (a) (division property) for any $a \in R$ and any nonzero $b \in R$ there exist $q, r \in R$ such that

$$a = bq + r,$$

with either $r = 0$ or $v(r) < v(b)$;

- (b) for any $a, b \in R$ with $ab \neq 0$, we have

$$v(a) \leq v(ab).$$

The canonical example of a Euclidean domain is the integers $\mathbb{Z}$ with the absolute value $v(x) = |x|$ as the valuation, but another important example is the ring of polynomials in one variable over a field.

Definition 15.2.2. *Define a valuation on the ring $\mathbb{F}[x]$ of polynomials with coefficients in $\mathbb{F}$ by $v(p(x)) = \deg p(x)$, where the degree $\deg p(x)$ of $p(x) = \sum_{i=0}^n a_i x^i$ is the greatest integer i such that a_i is not zero. For convenience, also define $\deg(0) = -\infty$.*

⁵³Almost everything we do in this chapter with the ring $\mathbb{F}[x]$ will work just as well with $F[x]$, where F is any field—not just $\mathbb{R}$ and $\mathbb{C}$. For a review of fields see Appendix B.2.

Proposition 15.2.3. *For any $a, b \in \mathbb{F}[x]$ the degree satisfies*

- (i) $\deg(ab) = \deg(a) + \deg(b)$,
- (ii) $\deg(a + b) \leq \max(\deg(a), \deg(b))$.

Proof. First, observe that $\mathbb{F}$ has no zero divisors; that is, if $\alpha, \beta \in \mathbb{F}$ are both nonzero, then the product $\alpha\beta$ is also nonzero. To see this, assume that $\alpha\beta = 0$. Since $\alpha \neq 0$, it has an inverse, so $\beta = (\alpha^{-1}\alpha)\beta = \alpha^{-1}(\alpha\beta) = \alpha^{-1} \cdot 0 = 0$, a contradiction. So the product of any two nonzero elements is not zero.

For (i), if a and b are nonzero, then writing out the polynomials $a = a_0 + a_1x + \cdots + a_mx^m$ and $b = b_0 + b_1x + \cdots + b_nx^n$, with b_n and a_m both nonzero, so that $\deg(a) = m$ and $\deg(b) = n$, we have

$$a \cdot b = a_0b_0 + (a_0b_1 + b_0a_1)x + \cdots + (a_{m-1}b_n + b_{n-1}a_m)x^{n+m-1} + a_mb_nx^{n+m}.$$

Since $a_mb_n \neq 0$, we have $\deg(ab) = n + m = \deg(a) + \deg(b)$.

For (ii), assume without loss of generality that $m \leq n$. If $m < n$, we have

$$a + b = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_m + b_m)x^m + b_{m+1}x^{m+1} + \cdots + b_nx^n,$$

and so in this case $\deg(a + b) = m = \max(\deg(a), \deg(b))$.

If $m = n$, we have

$$a + b = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_m + b_m)x^m.$$

So, if $a_m + b_m \neq 0$, we have $\deg(a + b) = \deg(a) = \deg(b)$, and if $a_m + b_m = 0$, we have $\deg(a + b) < \deg(a) = \deg(b)$.

The result also holds if one or both of the polynomials is 0. $\square$

Theorem 15.2.4. *The ring $\mathbb{F}[x]$ is a Euclidean domain with valuation given by the degree of the polynomial $v(p(x)) = \deg p(x)$.*

Proof. First observe that $\mathbb{F}[x]$ has no zero divisors because if $a, b \in \mathbb{F}[x]$ are both nonzero, then $\deg(ab) = \deg(a) + \deg(b) > -\infty = \deg(0)$.

Given $a \in \mathbb{F}[x]$ and any nonzero $b \in \mathbb{F}[x]$, let $S = \{a - bq \mid q \in \mathbb{F}[x]\}$. If $0 \in S$, then the proof is done, since there is a q such that $a = bq$. If $0 \notin S$, let $D = \{\deg(a - bq) \mid (a - bq) \in S\} \subset \mathbb{N}$ be the set of degrees of elements of S . By the well-ordering axiom of natural numbers (Axiom A.3.3), D has a least element d . Let r be some element of S with $\deg(r) = d$, so $r = a - bq$ for some $q \in \mathbb{F}[x]$.

We claim $d = \deg(r) < \deg(b)$. If not, then $b = b_0 + b_1x + \cdots + b_nx^n$ and $r = r_0 + r_1x + \cdots + r_mx^m$ with $m \geq n$. But now let $r' = r - b \frac{r_m}{b_n}x^{m-n} \in S$, so that the degree- m term of r' cancels. We have $\deg(r') < \deg(r)$ —a contradiction. $\square$

The degree function also gives a good way to characterize the invertible elements of $\mathbb{F}[x]$.

Proposition 15.2.5. *An element $f \in \mathbb{F}[x]$ is invertible if and only if $\deg(f) = 0$.*

Proof. For any invertible $f \in \mathbb{F}[x]$, we have $0 = \deg(1) = \deg(ff^{-1}) = \deg(f) + \deg(f^{-1})$, which implies that $\deg(f) = 0$. Conversely, if $\deg(f) = 0$, then $f = a_0 \in \mathbb{F}$ and $f \neq 0$, so $f^{-1} = a_0^{-1} \in \mathbb{F} \subset \mathbb{F}[x]$. $\square$

15.2.2 The Euclidean Algorithm in a Euclidean Domain

In this section we extend the familiar idea of divisibility from the integers to Euclidean domains. We also show that the very powerful Euclidean algorithm for finding the greatest common divisor works in any Euclidean domain.

We first need a quick discussion of divisibility properties in a Euclidean domain.

Definition 15.2.6. Let R be a Euclidean domain and let $a, b \in R$. If $b = ac$ for some $c \in R$, we say a divides b and write $a|b$.

Proposition 15.2.7. Any ideal I in a Euclidean domain R can be generated by a single element; that is, there exists some element $d \in I$ such that $(d) = I$. Moreover, $v(d)$ is least among all nonnegative integers of the form $v(i)$ for $i \in I \setminus 0$.

Proof. Let $S = \{n \in \mathbb{Z} \mid \exists i \in I \setminus \{0\}, n = v(i)\}$ be the image of the valuation map $v : I \setminus \{0\} \rightarrow \mathbb{N}$. By the well-ordering axiom of the integers, the set S must have a least element u , and there must be some element $d \in I \setminus 0$ such that $v(d) = u$.

Let $(d) \subset R$ be the ideal generated by d . Since $d \in I$, we have $(d) \subset I$. Given any $i \in I$, apply the division property (Definition 15.2.1(iv)(a)) to get $i = qd + r$ for some r with $v(r) < v(d) = u$. Since $d, i \in I$, we must have $r \in I$, but $v(r) < u$ contradicts the minimality of u unless $r = 0$. Therefore, $i = qd$ and $i \in (d)$. This proves that $I = (d)$. $\square$

Proposition 15.2.8. If R is a Euclidean domain and $a, b \in R$, there exists an element $d \in R$ that satisfies the following properties:

- (i) The element d can be written as $(ax + by)$ for $x, y \in R$, and

$$v(d) = \min\{v(ax' + by') : s', t' \in R\}.$$

- (ii) The element d divides both a and b , and any element d' with $d'|a$ and $d'|b$ must satisfy $d'|d$.

Moreover, any element that satisfies one of these properties must satisfy the other property, and if elements d, e satisfy these properties, then $d = ue$, where u is an invertible element of R .

Proof. By the previous proposition, the ideal (a, b) is generated by a single element d , and because $d \in (a, b)$, we have $d = ax + by$. In fact, an element c can be written as $c = ax' + by'$ for some $x', y' \in R$ if and only if $c \in (a, b)$. Moreover, any $c \in (a, b)$ must be divisible by d ; hence, by the multiplicative property of valuations (Definition 15.2.1(iv)(b)), we have $v(c) \geq v(d)$; so $v(d)$ is least in the set of all nonnegative integers of the form $v(ax' + by')$.

Since $(d) = (a, b)$ we have $d|a$ and $d|b$; conversely, given any d' with $d'|a$ and $d'|b$, we immediately have $d'|(ax + by) = d$. Now given any element $d \in R$ of the form $ax + by$ such that $v(d)$ is least among the nonnegative integers of the form $v(ax' + by')$, then the previous proposition shows that $(d) = (a, b)$; hence the second property must hold.

Conversely, if d is an element such that the second property holds, then by Proposition 15.2.7 we have $(a, b) = (e)$ for some $e \in (a, b)$, and the second property must also hold for e . Thus we have $d|e$ and $e|d$, so $e = ud$ and $d = u'e$. Therefore $e = u(u'e) = (uu')e$, so $e(1 - uu') = 0$. Since e is not zero and since R has no zero divisors, we must have $1 - uu' = 0$, or $1 = uu'$. So u is an invertible element of R . Moreover, $(d) = (e) = (a, b)$, so the first property must also hold. $\square$

Remark 15.2.9. The element d in the previous proposition is not necessarily unique. In fact, given any invertible element $\alpha \in R$ the element αd also satisfies all the conditions of the proposition. When $R = \mathbb{Z}$ we usually rescale to make the element d be positive, and we denote the resulting, rescaled element $\gcd(a, b)$. Similarly, when $R = \mathbb{F}[x]$, we usually rescale d so that it is monic—we also call this rescaled element $\gcd(a, b)$.

But for a general Euclidean domain, there is no canonical way to identify a unique element to call *the* gcd. Instead, we say that any element d is a gcd of a and b if it satisfies the conditions of the previous proposition.

Definition 15.2.10. In a Euclidean domain, for any tuple of elements $a_1, \dots, a_n$, we say that $d \in R$ is a gcd of $a_1, \dots, a_n$ if the ideal (d) equals the ideal $(a_1, \dots, a_n)$ of R . Elements $a_1, \dots, a_n$ are relatively prime if $(a_1, \dots, a_n) = R$. Since $R = (1) = (u)$ for any invertible element $u \in R$, we often write $\gcd(a_1, \dots, a_n) = 1$ to indicate that $a_1, \dots, a_n$ are relatively prime.

Remark 15.2.11. The previous discussion shows that two elements a and b in a Euclidean domain R are relatively prime if and only if the identity element 1 can be written $1 = as + bt$ for some $s, t \in R$.

Theorem 15.2.12 (Euclidean Algorithm). Let R be a Euclidean domain R , and let $a, b \in R$. Define q_0 and r_0 as in the division property (Definition 15.2.1(iv)(a)):

$$a = bq_0 + r_0.$$

Apply the division property to b and r to get q_1, r_1

$$b = r_0q_1 + r_1.$$

Repeating the process, eventually the remainder will be zero:

$$\begin{aligned} a &= bq_0 + r_0, \\ b &= r_0q_1 + r_1, \\ r_0 &= r_1q_2 + r_2, \\ r_1 &= r_2q_3 + r_3, \\ &\vdots \end{aligned}$$

$$\begin{aligned}r_{n-2} &= r_{n-1}q_n + r_n, \\r_{n-1} &= r_nq_{n+1} + 0.\end{aligned}$$

The result r_n is a greatest common divisor of a and b .

Proof. The algorithm terminates in no more than $v(b) + 1$ steps, because at each stage we have that $0 \leq v(r_k) < v(r_{k-1})$, so we have a sequence $v(b) > v(r_0) > v(r_1) > \dots \geq 0$ that decreases at each step until it reaches 0.

Since r_n divides r_{n-1} , and $r_{n-2} = r_{n-1}q_n + r_n$, we have that r_n divides r_{n-2} . Repeating the argument for the previous stages shows that r_n divides r_{n-3} and each r_k for $k = (n-3), (n-4), \dots, 1, 0$. Hence r_n divides both b and a . Conversely, given any c that divides a and b , the first equation $a = bq_0 + r_0$ shows that $c|r_0$. Repeating for each step gives $c|r_k$ for all k . Hence $c|r_n$. This implies that $r_n = \gcd(a, b)$. $\square$

Example 15.2.13. Let $a = x^4 - 1$ and $b = 2x^3 - 12x^2 + 22x - 12$ in the ring $\mathbb{C}[x]$. We apply the Euclidean algorithm to a and b to compute the greatest common divisor:

$$\begin{aligned}\underbrace{x^4 - 1}_a &= \underbrace{\left(\frac{x}{2} + 3\right)}_{q_0} \underbrace{(2x^3 - 12x^2 + 22x - 12)}_b \\&\quad + \underbrace{(25x^2 - 60x + 35)}_{r_0} \\ \underbrace{(2x^3 - 12x^2 + 22x - 12)}_b &= \underbrace{\left(\frac{2}{25}x - \frac{36}{125}\right)}_{q_1} \underbrace{(25x^2 - 60x + 35)}_{r_0} + \underbrace{\left(\frac{48}{25}x - \frac{48}{25}\right)}_{r_1} \\ \underbrace{(25x^2 - 60x + 35)}_{r_0} &= \underbrace{\left(\frac{625}{48}x - \frac{875}{48}\right)}_{q_2} \underbrace{\left(\frac{48}{25}x - \frac{48}{25}\right)}_{r_1} + \underbrace{0}_{r_2}.\end{aligned}$$

Thus $r_1 = \left(\frac{48}{25}x - \frac{48}{25}\right)$ is the desired common divisor of a and b , and $\gcd(a, b) = (x - 1)$.

Remark 15.2.14 (Extended Euclidean Algorithm (EEA)). For any $a, b \in R$, Proposition 15.2.8 guarantees the element $\gcd(a, b)$ can be written as $as + bt$ for some $s, t \in R$. Knowing the actual values of s and t is useful in many applications.

In the Euclidean algorithm, we have $r_n = r_{n-2} - r_{n-1}q_{n-1}$, and $r_{n-1} = r_{n-3} - r_{n-2}q_{n-2}$, and so forth, up to $r_0 = a - bq_0$. Back substituting gives an explicit expression for $\gcd(a, b) = r_n$ as $as + bt$. This is called the *extended Euclidean algorithm (EEA)*.

Example 15.2.15. Applying the EEA to the results of Example 15.2.13 gives $r_1 = b - q_1 r_0 = b - q_1(a - q_0 b) = (1 + q_0)b - q_1 a$, so $s = -q_1$ and $t = (1 + q_0)$ gives $r_1 = as + bt$.

15.3 The Fundamental Theorem of Arithmetic

The fundamental theorem of arithmetic states that every integer is a product of prime integers, and that decomposition is unique (except for choices of signs on the primes). The analogous result for Euclidean domains is a powerful tool and plays a role in applications of the Chinese remainder theorem and in building partial fraction decompositions.

We begin with the observation that even though a Euclidean domain does not have a multiplicative inverse for every element, we can still cancel nonzero factors.

Proposition 15.3.1 (Cancellation). *In a Euclidean domain R , if $a, b, c \in R$ and $a \neq 0$, then $ab = ac$ implies $b = c$.*

Proof. If $ab = ac$, then $ab - ac = 0$ and so $a(b - c) = 0$. Since $a \neq 0$ and since R has no zero divisors, then we must have $(b - c) = 0$, and hence $b = c$. $\square$

Definition 15.3.2. *An element $a \in R$ is called prime if it is not zero and not invertible, and if whenever $a|bc$ for some $b, c \in R$, then $a|b$ or $a|c$. An element $a \in R$ is called irreducible if it is not zero and not invertible, and whenever $a = bc$, then either b or c is invertible in R .*

Remark 15.3.3. The traditional definition of a prime integer is our definition of irreducible, since the only invertible elements in $\mathbb{Z}$ are 1 and -1 . In a general ring, prime and irreducible are not the same thing, but we show below that they are equivalent in a Euclidean domain.

Proposition 15.3.4. *If two elements a, b in a Euclidean domain R are relatively prime and $c \in R$, then $a|bc$ implies $a|c$.*

Proof. Since $a|bc$, we have $bc = ad$ for some $d \in R$. Since a, b are relatively prime, we have $\gcd(a, b) = 1$, so $1 = ax + by$ for some $x, y \in R$. Thus $c = axc + byc = axc + yad = a(xc + yd)$, which implies that $a|c$. $\square$

Corollary 15.3.5. *Let R be a Euclidean domain, and let $m_1, m_2, \dots, m_n \in R$ be pairwise relatively prime; that is, for every $i \neq j$ we have $\gcd(m_i, m_j) = 1$. If $x \in R$ is such that m_i divides x for every $i \in \{1, \dots, n\}$, then $\prod_{i=1}^n m_i$ divides x .*

Proof. The proof is by induction on n . If $n = 1$, the result is immediate. If $n > 1$, then $m_n|x$ implies that $x = m_n y$ for some $y \in R$, and $m_{n-1}|x$ implies $m_{n-1}|m_n y$. By Proposition 15.3.4, this means that $m_{n-1}|y$, so $y = m_{n-1} z$ for some $z \in R$, and

we have that $x = m_n m_{n-1} z$ is divisible by $m'_{n-1} = m_n m_{n-1}$. Now make a new list $m_1, \dots, m_{n-2}, m'_{n-1}$. This list has length $n-1$, and the elements in it are pairwise relatively prime, and each divides x . By the induction hypothesis x is divisible by $m_1 m_2 \dots m_{n-2} m'_{n-1} = \prod_{i=1}^n m_i$. $\square$

Corollary 15.3.6. *If $I_1, I_2, \dots, I_n$ are ideals in a Euclidean domain such that $I_j = (m_j)$ and the generators m_i are pairwise relatively prime, then the ideal $(m_1 m_2 \dots m_n)$ generated by the product $m_1 m_2 \dots m_n$ satisfies*

$$(m_1 m_2 \dots m_n) = \bigcap_{i=1}^n I_i = \prod_{i=1}^n I_i.$$

Proof. Clearly $m_1 m_2 \dots m_n \in \prod_{i=1}^n I_i \subset \bigcap_{i=1}^n I_i$, and hence $(m_1 m_2 \dots m_n) \subset \prod_{i=1}^n I_i \subset \bigcap_{i=1}^n I_i$. Conversely, for any $x \in \bigcap_{i=1}^n I_i$, we have $x \in I_i$ for every $i \in \{1, \dots, n\}$, and hence $m_i | x$ for every i . By the Corollary 15.3.5, x is divisible by $m_1 m_2 \dots m_n$, and hence $x \in (m_1 m_2 \dots m_n)$. $\square$

Theorem 15.3.7. *An element in a Euclidean domain is prime if and only if it is irreducible.*

Proof. If $a \in R$ is prime, and if $a = bc$, then a divides bc , so $a|b$ or $a|c$. Without loss of generality, assume $a|b$. We have $b = ax$ for some $x \in R$, and $a \cdot 1 = bc = axc$. The cancellation property (Proposition 15.3.1) gives $1 = xc$, which implies that both x and c are invertible.

Conversely, assume an irreducible element a divides bc for some $b, c \in R$. If $\gcd(a, b) = 1$, then Proposition 15.3.4 gives $a|c$, as required.

If $\gcd(a, b) = d$ is not invertible, then $a = dx$ for some $x \in R$. By irreducibility of a , the element x is invertible in R . Thus, the element $ax^{-1} = d$ divides b , and hence a divides b . $\square$

Proposition 15.3.8. *Any element f of $\mathbb{F}[x]$ that is of degree 1 must be irreducible, and hence prime.*

Proof. The proof is Exercise 15.22. $\square$

Remark 15.3.9. Not every prime element in $\mathbb{F}[x]$ has degree 1. In particular, the element $x^2 + 1 \in \mathbb{R}[x]$ is irreducible in $\mathbb{R}[x]$ but has degree 2.

Before proving the main result of this section, we need one more result about the multiplicative property of the valuation in a Euclidean domain.

Proposition 15.3.10. *For every $s, t \in R$, we have $v(s) = v(st)$ if and only if t is invertible.*

Proof. If t is invertible, then $v(s) \leq v(st) \leq v(stt^{-1}) = v(s)$, so $v(s) = v(st)$. Conversely, if $v(s) = v(st)$, then $s = (st)q + r$ for some $r \in R$ with $v(r) < v(st)$. But $r = s - (st)q = s(1 - tq)$; so $v(s) \leq v(s(1 - tq)) = v(r)$ unless $1 - tq = 0$. Hence t is invertible. $\square$

Theorem 15.3.11 (Fundamental Theorem of Arithmetic for Euclidean Domains). *Any nonzero element a in a Euclidean domain R can be written as a product of primes times some invertible element. Moreover, if $a = \alpha p_1 p_2 \cdots p_n$ and $a = \beta q_1 q_2 \cdots q_m$ are two such decompositions, with α, β invertible and all p_i and q_i prime, then $m = n$. Reordering if necessary gives $p_i = u_i q_i$, where each u_i is an invertible element in R .*

Proof. Let S be the set of all nonzero, noninvertible elements of R that cannot be written as an invertible element times a product of primes. The set $V = \{v(s) | s \in S\}$ is a subset of $\mathbb{N}$ and hence has a smallest element v_0 . Let $s \in S$ be an element with $v(s) = v_0$. Since s is not a product of primes, it is not prime, and hence not irreducible. Therefore, there exist $a, b \in R$ such that $ab = s$ and a and b are not invertible. Proposition 15.3.10 implies $v(s) > v(a)$ and $v(s) > v(b)$, and hence a and b can be written in the desired form. But $s = ab$ implies that s can also be written in the desired form. Therefore, V and S must both be empty.

To prove uniqueness, assume that $a = \alpha p_1 p_2 \cdots p_n$ and $a = \beta q_1 q_2 \cdots q_m$ are two decompositions that do not satisfy the conclusion of the theorem, that is, either $n \neq m$ or $n = m$, but there is no rearrangement of the q_j such that every $q_i = u_i p_i$ for invertible u_i . Assume further that n is the smallest integer for which such a counterexample exists.

If $n = 0$, then $a = \alpha = \beta q_1 \cdots q_m$, so a is invertible. Thus, $q_i | a$ implies that q_i is also invertible for every i , and hence q_i is not a prime. So we may assume $n > 0$.

Since p_n is prime, it must divide q_i for some $i \in \{1, \dots, m\}$. Rearrange the q_j so that p_n divides q_m . Thus $p_n u_n = q_m$ for some u_n . But since q_m is prime, it is irreducible. Since p_n is not invertible, u_n must be invertible.

Now divide p_n out of both sides of the equation $\alpha p_1 p_2 \cdots p_n = \beta q_1 q_2 \cdots q_m$ to get $\alpha p_1 p_2 \cdots p_{n-1} = \beta q_1 q_2 \cdots q_{m-1} u_n$. Redefining q_{m-1} to be $q_{m-1} u_n$ gives two new decompositions into primes $\alpha p_1 p_2 \cdots p_{n-1} = \beta q_1 q_2 \cdots q_{m-1}$, and by the minimality assumption on the original counterexample gives $m-1 = n-1$, and (after reordering) $p_i = u_i q_i$, where each u_i is an invertible element in R . But this proves that the supposed counterexample also satisfies the conclusion of the theorem. $\square$

Example 15.3.12.

- (i) The integer 1728 can be written as $3^3 2^6$ or $(-2)^3 2^3 (-3)^3$ and in many other ways, but the fundamental theorem of arithmetic says that, after rearranging, every prime factorization of 1728 must be of the form

$$\pm 3 \cdot \pm 3 \cdot \pm 3 \cdot \underbrace{\pm 2 \cdots \pm 2}_6.$$

(ii) The polynomial

$$p = x^9 - 20x^8 + 129x^7 - 257x^6 + 59x^5 - 797x^4 - 265x^3 - 903x^2 - 196x - 343$$

can be factored in $\mathbb{R}[x]$ as

$$(x^2 + 1)^2(x - 7)^3(x^2 + x + 1),$$

and it can be shown that in $\mathbb{R}[x]$ the polynomials $x^2 + 1$ and $x^2 + x + 1$ are both prime, as is every linear polynomial. Therefore, this is a prime factorization of p . The fundamental theorem of arithmetic shows that any other factorization of p must (after rearranging) be of the form

$$p = a(x^2 + 1) \cdot b(x^2 + 1) \cdot c(x - 7) \cdot d(x - 7) \cdot e(x - 7) \cdot f(x^2 + x + 1),$$

where a, b, c, d, e, f are invertible and, hence, are in $\mathbb{R}$.

Unexample 15.3.13. Exercise 15.23 shows that the fundamental theorem of arithmetic does not hold in the ring

$$\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\} \subset \mathbb{C},$$

because there are two distinct prime factorizations of 9. Therefore $\mathbb{Z}[\sqrt{-5}]$ is not a Euclidean domain.

Remark 15.3.14. This theorem says that every element is almost unique as a product of primes. In both cases $R = \mathbb{F}[x]$ and $R = \mathbb{Z}$ we can make the uniqueness more explicit.

If $R = \mathbb{Z}$, then the only invertible elements are 1 and -1 ; if $a > 0$ we may require all primes to be positive. We have $u_i = 1$ for all i , and the decomposition is completely unique (up to reordering).

If $R = \mathbb{F}[x]$, then the invertible elements are precisely the elements of $\mathbb{F}$ (corresponding to degree-zero polynomials). If both a and all the primes have their leading (top-degree) coefficient equal to 1, then we can again assume that all the u_i are 1, and the decomposition is completely unique (up to reordering).

Theorem 15.3.15 (Fundamental Theorem of Algebra, alternative form).

All primes in $\mathbb{C}[x]$ have degree 1, and hence every polynomial f in $\mathbb{C}[x]$ of degree n can be factored uniquely (up to rearrangement of the factors) as

$$f(x) = c \prod_{i=1}^n (x - \lambda_i),$$

where $c, \lambda_1, \dots, \lambda_n \in \mathbb{C}$.

Proof. Assume, by way of contradiction, that $p(x)$ is prime in $\mathbb{C}[x]$ and has degree $n > 1$. By the fundamental theorem of algebra (Theorem 11.5.4), $p(x)$ has at least one root, which we denote λ . Dividing $p(x)$ by $(x - \lambda)$ gives

$$p(x) = (x - \lambda)q(x) + r(x),$$

where $q(x)$ has degree $n - 1$, and where $r(x)$ has degree less than $\deg(x - \lambda)$; hence r is constant. Moreover, $0 = p(\lambda) = 0 + r$, and hence $r = 0$. Therefore, $(x - \lambda)$ divides $p(x)$, and p is not prime. $\square$

15.4 Homomorphisms

A linear transformation is the right sort of map for vector spaces because it preserves all the key properties of a vector space—vector addition and scalar multiplication. Similarly, a *ring homomorphism* is the right sort of map for rings because it preserves the key properties of a ring—addition and multiplication.

Just as with vector spaces, kernels and ranges (images) are the key to understanding ring homomorphisms, and invertible homomorphisms (isomorphisms) allow us to identify which rings are “the same.”

15.4.1 Homomorphisms

Definition 15.4.1. A map $f : R \rightarrow S$ between rings is a homomorphism from R into S if

$$f(xy) = f(x)f(y) \quad \text{and} \quad f(x + y) = f(x) + f(y) \quad (15.2)$$

for all $x, y \in R$.

The next proposition is immediate, and in fact it may even seem more difficult than checking the definition directly, but it makes many proofs much cleaner; it also makes the similarity to linear transformations clearer.

Proposition 15.4.2. To check that a map $f : R \rightarrow S$ of rings is a homomorphism, it suffices to show that for every $a, b, x, y \in R$ we have $f(ax + by) = f(a)f(x) + f(b)f(y)$.

Example 15.4.3.

- (i) For any $n \in \mathbb{Z}$ the map $\mathbb{Z} \rightarrow \mathbb{Z}_n$ given by $x \mapsto [x]_n$ is a ring homomorphism.
- (ii) For any interval $[a, b] \subset \mathbb{R}$, the map $\mathbb{F}[x] \rightarrow C([a, b]; \mathbb{F})$ given by sending a polynomial $f(x) \in \mathbb{F}[x]$ to the function on $[a, b]$ defined by f is a homomorphism of rings.

- (iii) For any $p \in \mathbb{F}$, the evaluation map $e_p : \mathbb{F}[x] \rightarrow \mathbb{F}$ defined by $f(x) \mapsto f(p)$ is a homomorphism of rings.
- (iv) For any $A \in M_n(\mathbb{F})$, the evaluation map $e_A : \mathbb{F}[x] \rightarrow \mathbb{F}[A] \subset M_n(\mathbb{F})$ defined by $f(x) \mapsto f(A)$ is a homomorphism of rings.

Unexample 15.4.4.

- (i) For any $n \in \mathbb{Z}^+$, the map $C^n((a, b); \mathbb{F}) \rightarrow C^{n-1}((a, b); \mathbb{F})$ defined by $f(x) \mapsto \frac{df(x)}{dx}$ is not a homomorphism—it preserves addition

$$\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx},$$

but it does not preserve multiplication

$$\frac{d(f \cdot g)}{dx} = g \frac{df}{dx} + f \frac{dg}{dx} \neq \frac{df}{dx} \cdot \frac{dg}{dx}.$$

- (ii) The map $\phi : \mathbb{F} \rightarrow \mathbb{F}$ given by $\phi(x) = x^2$ is not a homomorphism because $\phi(x+y)$ is not always (in fact, almost never) equal to $x^2 + y^2 = \phi(x) + \phi(y)$.

Proposition 15.4.5. *A homomorphism of rings $f : R \rightarrow S$ maps the additive identity $0 \in R$ to the additive identity $0 \in S$.*

Proof. The proof is identical to its linear counterpart, Proposition 2.1.5(ii). $\square$

Proposition 15.4.6. *If $f : R \rightarrow S$ and $g : S \rightarrow U$ are homomorphisms, then the composition $g \circ f : R \rightarrow U$ is also a homomorphism.*

Proof. Given $a, b, x, y \in R$ we have

$$\begin{aligned} (g \circ f)(ax + by) &= g(f(a)f(x) + f(b)f(y)) \\ &= g(f(a))g(f(x)) + g(f(b))g(f(y)) \\ &= (g \circ f)(a)(g \circ f)(x) + (g \circ f)(b)(g \circ f)(y). \quad \square \end{aligned}$$

Remark 15.4.7. Unlike its vector space analogue, a homomorphism does not necessarily map ideals into ideals. But if the homomorphism is surjective, then it does preserve ideals.

15.4.2 The Kernel and Image

Definition 15.4.8. Let R and S be rings. The kernel of a homomorphism $f : R \rightarrow S$ is the set $\mathcal{N}(f) = \{x \in R \mid f(x) = 0\}$. The image (or range) of f is the set $\text{Im } f = \mathcal{R}(f) = \{f(x) \in S \mid x \in R\}$.

Proposition 15.4.9. For any homomorphism f

- (i) the set $\mathcal{N}(f)$ is an ideal of R ;
- (ii) the set $\text{Im } f$ is a subring of S (but not necessarily an ideal).

Proof. Note that $\mathcal{N}(f)$ and $\text{Im } f$ are both nonempty since $f(0) = 0$.

- (i) Let $x_1, x_2 \in \mathcal{N}(f)$ and $a, b \in R$. By definition, $f(x_1) = f(x_2) = 0$; hence $f(ax_1 + bx_2) = f(a)f(x_1) + f(b)f(x_2) = f(a)0 + f(b)0 = 0$. Therefore $ax_1 + bx_2 \in \mathcal{N}(f)$.
- (ii) For any $s, t, y_1, y_2 \in \text{Im}(f)$ there exist $a, b, x_1, x_2 \in R$ such that $f(a) = s$, $f(b) = t$, and $f(x_i) = y_i$ for $i \in \{1, 2\}$. Thus we have $sy_1 + ty_2 = f(a)f(x_1) + f(b)f(x_2) = f(ax_1 + bx_2) \in \text{Im } f$. $\square$

Example 15.4.10. If $f : \mathbb{Z} \rightarrow \mathbb{Z}_2$ is given by $x \mapsto \llbracket x \rrbracket_2$, then $\mathcal{N}(f) = 2\mathbb{Z} = \mathbb{E}$ is the set of all even integers.

Example 15.4.11.  For any $A \in M_n(\mathbb{F})$, let $e_A : \mathbb{F}[x] \rightarrow \mathbb{F}[A] \subset M_n(\mathbb{F})$ be the evaluation homomorphism defined by $f(x) \mapsto f(A)$. The kernel $\mathcal{N}(e_A)$ consists of precisely those polynomials $p(x)$ such that $p(A) = 0$. In particular, $\mathcal{N}(e_A)$ contains the characteristic polynomial and the minimal polynomial. Since every ideal in $\mathbb{F}[x]$ is generated by any element of least degree (see Proposition 15.2.7), and since the minimal polynomial is defined to have the least degree of any element in $\mathbb{F}[x]$ that annihilates A , the ideal $\mathcal{N}(e_A)$ must be generated by the minimal polynomial.

Example 15.4.12.  Let $p \in \mathbb{F}$ be any point, and let $e_p : \mathbb{F}[x] \rightarrow \mathbb{F}$ be the evaluation homomorphism, given by $e_p(f) = f(p)$. If f is a constant polynomial $f = c$, then $e_p(f) = c$. This shows that e_p is surjective and $\text{Im}(e_p) = \mathbb{F}$. On the other hand, we have that $e_p(x - p) = 0$, so $x - p \in \mathcal{N}(e_p)$.

Since every ideal in $\mathbb{F}[x]$ is generated by a single element, we must have $\mathcal{N}(e_p) = (g)$ for some polynomial $g \in \mathbb{F}[x]$. In particular, we have $g|(x-p)$, but since $x-p$ is of degree 1, it is prime, and hence $g = (x-p)u$ for some invertible element u . Thus $(x-p) = (g) = \mathcal{N}(e_p)$.

For any $p \neq q \in \mathbb{F}$ the generator $x-p$ of $\mathcal{N}(e_p)$ and the generator $x-q$ of $\mathcal{N}(e_q)$ are both of degree 1, hence prime, by Proposition 15.3.8. Thus, by the fundamental theorem of arithmetic, their powers are relatively prime. This gives the following proposition.

Proposition 15.4.13. *If m, n are nonnegative integers and $p \neq q \in \mathbb{F}$ are arbitrary points of $\mathbb{F}$, then the polynomials $(x-p)^m$ and $(x-q)^n$ are relatively prime.*

15.4.3 Isomorphisms: Invertible Homomorphisms

Definition 15.4.14. *A homomorphism of rings $f : R \rightarrow S$ is called an isomorphism if it has an inverse that is also a homomorphism. An isomorphism from R to R is called an automorphism.*

As in the case of linear transformations of vector spaces, if a ring homomorphism has an inverse, then the inverse must also be a homomorphism.

Proposition 15.4.15. *If a homomorphism is bijective, then the inverse function is also a homomorphism.*

Proof. Assume that $f : R \rightarrow S$ is a bijective homomorphism with inverse function f^{-1} . Given $y_1, y_2, s, t \in S$ there exists $x_1, x_2, a, b \in R$ such that $f(x_i) = y_i$, $f(a) = s$, and $f(b) = t$. Thus we have

$$\begin{aligned} f^{-1}(sy_1 + ty_2) &= f^{-1}(f(a)f(x_1) + f(b)f(x_2)) \\ &= f^{-1}(f(ax_1 + bx_2)) \\ &= ax_1 + bx_2 \\ &= f^{-1}(s)f^{-1}(y_1) + f^{-1}(t)f^{-1}(y_2). \end{aligned}$$

Therefore f^{-1} is a homomorphism. $\square$

Corollary 15.4.16. *A homomorphism is an isomorphism if and only if it is bijective.*

Example 15.4.17.

- (i) Let $R = \{\text{True}, \text{False}\}$ be the Boolean ring of Example 15.1.3(vii), and let S be the ring $\mathbb{Z}_2$. The map $\text{False} \mapsto 0$ and $\text{True} \mapsto 1$ is an isomorphism.

(ii) Let $S \subset M_2(\mathbb{R})$ be the set of all 2×2 real matrices of the form

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix}, \quad \text{with } a, b \in \mathbb{R}.$$

It is not hard to see that S with matrix addition and multiplication is a ring. In fact, the map $\phi : S \rightarrow \mathbb{C}$ given by

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mapsto a + bi$$

is an isomorphism of rings. The proof is Exercise 15.29.

Definition 15.4.18. If there exists an isomorphism $L : R \rightarrow S$, we say that R is isomorphic to S , and we denote this by $R \cong S$.

Theorem 15.4.19. The relation $\cong$ is an equivalence relation on the collection of all rings.

Proof. The proof is identical to its counterpart for invertible linear transformations (Theorem 2.2.12). $\square$

15.4.4 Cartesian Products

Proposition 15.4.20. Let $\{R_1, R_2, \dots, R_n\}$ be a collection of rings. The Cartesian product

$$R = \prod_{i=1}^n R_i = R_1 \times R_2 \times \cdots \times R_n = \{(a_1, a_2, \dots, a_n) \mid a_i \in R_i\}$$

forms a ring with additive identity $(0, 0, \dots, 0)$ and with componentwise addition and multiplication. That is, addition and multiplication are given by

$$(i) \quad (a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n),$$

$$(ii) \quad (a_1, a_2, \dots, a_n) \cdot (b_1, b_2, \dots, b_n) = (a_1 \cdot b_1, a_2 \cdot b_2, \dots, a_n \cdot b_n)$$

for all $(a_1, a_2, \dots, a_n), (b_1, b_2, \dots, b_n) \in R$.

Proof. The proof is Exercise 15.30. $\square$

Example 15.4.21. The ring $\mathbb{Z}_3 \times \mathbb{Z}_2$ consists of 6 elements:

$$\begin{aligned} & ([0]_3, [0]_2), \quad ([1]_3, [0]_2), \quad ([2]_3, [0]_2), \\ & ([0]_3, [1]_2), \quad ([1]_3, [1]_2), \quad ([2]_3, [1]_2). \end{aligned}$$

Adding the element $([1]_3, [1]_2)$ to itself repeatedly gives

$$\begin{aligned} ([1]_3, [1]_2) + ([1]_3, [1]_2) &= ([2]_3, [0]_2), \\ ([2]_3, [0]_2) + ([1]_3, [1]_2) &= ([0]_3, [1]_2), \\ ([0]_3, [1]_2) + ([1]_3, [1]_2) &= ([1]_3, [0]_2), \\ ([1]_3, [0]_2) + ([1]_3, [1]_2) &= ([2]_3, [1]_2), \\ ([2]_3, [0]_2) + ([1]_3, [1]_2) &= ([0]_3, [0]_2). \end{aligned}$$

But multiplying $([1]_3, [1]_2)$ times any $(x, y) \in \mathbb{Z}_3 \times \mathbb{Z}_2$ gives (x, y) again; that is, $([1]_3, [1]_2)$ is the multiplicative identity element for this ring.

There are two obvious homomorphisms from the ring $\mathbb{Z}_3 \times \mathbb{Z}_2$, namely, the first projection $\mathbb{Z}_3 \times \mathbb{Z}_2 \rightarrow \mathbb{Z}_3$, defined by $([a], [b]) \mapsto [a]$, and the second projection $\mathbb{Z}_3 \times \mathbb{Z}_2 \rightarrow \mathbb{Z}_2$, given by $([a], [b]) \mapsto [b]$.

Proposition 15.4.22. *Given a collection of rings $R_1, \dots, R_n$, and given any $i \in \{1, \dots, n\}$, the canonical projection $p_i : \prod_{j=1}^n R_j \rightarrow R_i$, given by $(x_1, \dots, x_n) \mapsto x_i$, is a homomorphism of rings.*

Proof. We check $p_i((x_1, \dots, x_n) + (y_1, \dots, y_n)) = p_i((x_1 + y_1, \dots, x_n + y_n)) = x_i + y_i = p_i(x_1, \dots, x_n) + p_i(y_1, \dots, y_n)$. The check for multiplication is similar. $\square$

Remark 15.4.23. It is straightforward to check that all the results proved so far about finite Cartesian products also hold for infinite Cartesian products.

15.4.5 * Universal Mapping Property

The next proposition says that giving a homomorphism $T \rightarrow \prod_{i=1}^n R_i$ is exactly equivalent to giving a collection of homomorphisms $T \rightarrow R_i$ for every i .

Proposition 15.4.24. *Let $\{R_i\}_{i=1}^n$ be a finite collection of rings, and let T be a ring. Any homomorphism $F : T \rightarrow \prod_{i=1}^n R_i$ determines a collection of homomorphisms $f_i : T \rightarrow R_i$ for each i . Conversely, given a collection $\{f_i : T \rightarrow R_i\}_{i=1}^n$ of homomorphisms, there is a uniquely determined homomorphism $F : T \rightarrow \prod_{i=1}^n R_i$ such that $p_i \circ F = f_i$ for every i ; that is, the following diagram commutes.⁵⁴*

$$\begin{array}{ccc} & & \prod_{i=1}^n R_i \\ & \nearrow \exists! F & \downarrow p_i \\ T & \xrightarrow{f_i} & R_i \end{array}$$

⁵⁴See Proposition A.2.15 for a set-theoretic version of this proposition.

Proof. If $F : T \rightarrow \prod_{i=1}^n R_i$ is a homomorphism, then each $f_i = p_i \circ F$ is also a homomorphism, because the composition of homomorphisms is a homomorphism.

Conversely, given functions $f_i : T \rightarrow R_i$ for every i , let F be given by $F(t) = (f_1(t), \dots, f_n(t))$. It is immediate that the composition $p_i \circ F$ satisfies $p_i \circ F(t) = f_i(t)$ and that it is the unique function mapping T to $\prod_{i=1}^n R_i$ that satisfies this condition. It remains to check that F is a homomorphism.

We check the addition property of a homomorphism—the computation for multiplication is similar.

$$\begin{aligned} F(s+t) &= (f_1(s+t), \dots, f_n(s+t)) \\ &= (f_1(s) + f_1(t), \dots, f_n(s) + f_n(t)) = F(s) + F(t). \quad \square \end{aligned}$$

15.5 Quotients and the First Isomorphism Theorem

Just as one may construct a quotient of a vector space by a subspace, one may also construct a quotient of a ring by an ideal. Many of the results about isomorphisms of vector spaces and quotients carry over almost exactly to rings. Throughout this section we omit the ring-theoretic proofs of any results that are similar to their vector-space counterparts and only include those that differ in some meaningful way from the vector-space version.

Definition 15.5.1. Let I be an ideal of R . We say that x is equivalent to y modulo I if $x - y \in I$; this is denoted $x \equiv y \pmod{I}$. Sometimes, instead of equivalent modulo I , we say congruent modulo I .

Proposition 15.5.2. Let I be an ideal of R . The relation $\equiv \pmod{I}$ is an equivalence relation.

Definition 15.5.3. The equivalence classes $[y] = \{y \mid y \equiv x \pmod{I}\}$ partition R , and each equivalence class $[y]$ is a translate of I , that is,

$$[y] = y + I.$$

We call these equivalence classes cosets of I and write them as either $y + I$ or $[y]_I$. If the ideal I is clear from context, we often write $[y]$ without the I .

Because they are equivalence classes, any two cosets are either identical or disjoint. That is, if $(x+I) \cap (y+I) = [x] \cap [y] \neq \emptyset$, then $x+I = [x] = [y] = y+I$.

Definition 15.5.4. Let I be an ideal of R . The set $\{[x] \mid x \in R\}$ of all cosets of I in R is denoted R/I and is called the quotient of the ring R by the ideal I .

Proposition 15.5.5. Let $I = (b)$ be any ideal in a Euclidean domain R . The quotient R/I has a bijection to the set $S = \{s \in R \mid v(s) < v(b)\}$.

Proof. Given any $[a] \in R/I$, the division property gives $a = bq + r$ with $v(r) < v(b)$, so $r \in S$. Define a map $\phi : R/I \rightarrow S$ by $[a]_I \mapsto r$. First we show this map ϕ is well defined. Given any other a' with $[a'] = [a]$, we have $a' - a \in I = (b)$; so there

exists an $x \in R$ with $a' = a + xb = (q + x)b + r$. This implies $\phi(a') = \phi(a)$, and hence ϕ is well defined.

Define another map $\psi : S \rightarrow R/I$ by $s \mapsto [s]_I$. For any $s \in S$, we have $s = 0b + s$, so we get $\phi \circ \psi(s) = s$, and conversely, for any $[a]_I \in R/I$, with $a = bq + r$, we have $\psi \circ \phi(a) = [r]_I = [a]_I$, so ϕ and ψ are inverses, and hence bijections. $\square$

Remark 15.5.6. The previous proposition means that we can write the set $\mathbb{Z}/(n)$ as the set $\{[0], [1], [2], \dots, [n-1]\}$, and, similarly, any coset in $\mathbb{F}[x]/(f)$ can be written uniquely as $[r]$ for some r of degree less than $\deg(f)$.

The set of all cosets of I has a natural addition and multiplication that makes it into a ring.

Lemma 15.5.7. *Let I be an ideal of R . The operations $\boxplus : R/I \times R/I \rightarrow R/I$ and $\boxdot : R/I \times R/I \rightarrow R/I$ given by*

$$(i) \quad (x + I) \boxplus (y + I) = (x + y) + I \text{ (or } [x] \boxplus [y] = [x + y]) \text{ and}$$

$$(ii) \quad (x + I) \boxdot (y + I) = (xy) + I \text{ (or } [x] \boxdot [y] = [xy])$$

are well defined for all $x, y \in R$.

Proposition 15.5.8. *Let I be an ideal of R . The quotient R/I is a ring, when endowed with the operations $\boxplus$ and $\boxdot$.*

Remark 15.5.9. It is common to use the symbol $+$ instead of $\boxplus$ and $\cdot$ instead of $\boxdot$. This can quickly lead to confusion unless you pay careful attention to distinguish the various meanings of these overloaded symbols.

Example 15.5.10.

(i) By Remark 15.5.6 the quotient $\mathbb{Z}/(n)$ is in bijective correspondence with the elements of $\{0, 1, \dots, n-1\}$, and hence to the elements of $\mathbb{Z}_n$. It is straightforward to check that the addition and multiplication in $\mathbb{Z}/(n)$ are the same as the addition and multiplication in $\mathbb{Z}_n$.

(ii) By Remark 15.5.6 every element of the quotient $\mathbb{R}[x]/(x^2 + 1)$ can be written uniquely as $[f]$ for some f of degree less than 2. That means it must be of the form $[ax + b]$ for some $a, b \in \mathbb{R}$. The operation $\boxplus$ is straightforward $[ax + b] \boxplus [cx + d] = [(a + c)x + (b + d)]$, but the operation $\boxdot$ is more subtle. Observe first that $[x^2] = [-1]$, which gives

$$[ax + b] \boxdot [cx + d] = [acx^2 + (ad + bc)x + bd] = [(ad + bc)x + (bd - ac)].$$

This is just like multiplication of complex numbers, which inspires a map $\zeta : \mathbb{R}[x]/I \rightarrow \mathbb{C}$ defined by $[ax + b] \mapsto ai + b$. Using the previous computation, it is straightforward to check that this is a homomorphism. It is also clearly both surjective and injective, hence an isomorphism.

Proposition 15.5.11. *If a ring is commutative, then for any ideal $I \subset R$ the quotient ring R/I is commutative.*

Proof. If $x+I$ and $y+I$ are any elements of R/I , then $(x+I) \square (y+I) = xy+I = yx+I = (y+I) \square (x+I)$. $\square$

Proposition 15.5.12. *Let I be an ideal of the ring R with quotient R/I . The mapping $\pi : R \rightarrow R/I$ defined by $\pi(x) = x+I$ is a surjective homomorphism. We call π the canonical epimorphism.*

Lemma 15.5.13. *A homomorphism f is injective if and only if $\mathcal{N}(f) = \{0\}$.*

Lemma 15.5.14. *Let R and S be rings, and let $f : R \rightarrow S$ be a homomorphism. Assume that I and J are ideals of R and S , respectively. If $f(I) \subset J$, then f induces a homomorphism $\bar{f} : R/I \rightarrow S/J$ defined by $\bar{f}(x+I) = f(x)+J$.*

Theorem 15.5.15 (First Isomorphism Theorem). *If R and S are rings and $f : R \rightarrow S$ is a homomorphism, then $R/\mathcal{N}(f) \cong \text{Im } f$, where $\text{Im } f$ is the image of f . In particular, if f is surjective, then $R/\mathcal{N}(f) \cong S$.*

Remark 15.5.16. Any homomorphism $f : R \rightarrow S$ gives the following commutative diagram:

$$\begin{array}{ccc}
 R & \xrightarrow{f} & S \\
 \pi \downarrow & \searrow \wr & \uparrow i \\
 R/\mathcal{N}(f) & \xrightarrow{\bar{f}} & \text{Im } f
 \end{array}$$

where the right-hand vertical map i is just the obvious inclusion of the image of f into S , and the bottom horizontal map $\bar{f}$ is an isomorphism.

This theorem is often used to prove that some quotient R/I is isomorphic to some other ring S . If you construct a surjective homomorphism $R \rightarrow S$ that has kernel equal to I , then the first isomorphism theorem then gives the desired isomorphism.

Example 15.5.17.

- (i) If R is a ring, then $R/R \cong \{0\}$. Define the homomorphism $L : R \rightarrow \{0\}$ by $f(x) = 0$ for all $x \in R$. The kernel of f is all of R , and the first isomorphism theorem gives the isomorphism.
- (ii) For any $p \in [a, b] \subset \mathbb{R}$ let $\mathfrak{m}_p = \{f \in C([a, b]; \mathbb{F}) \mid f(p) = 0\}$. We claim that $C([a, b]; \mathbb{F})/\mathfrak{m}_p \cong \mathbb{F}$. To see this, use the homomorphism $e_p : C([a, b]; \mathbb{F}) \rightarrow \mathbb{F}$ given by $e_p(f) = f(p)$. The map e_p is clearly

surjective, and the kernel of e_p is precisely $\mathfrak{m}_p$, so the first isomorphism theorem gives the desired isomorphism.

- (iii) Example 15.5.10(ii) shows that the ring of complex numbers $\mathbb{C}$ is isomorphic to $\mathbb{R}[x]/(x^2 + 1)$. The first isomorphism theorem gives another way to see this: define a map $\mathbb{R}[x] \rightarrow \mathbb{C}$ by $f(x) \mapsto f(i)$, where $i \in \mathbb{C}$ is the usual square root of -1 . It is easy to see that this map is a homomorphism. Its kernel is the set of all polynomials $f \in \mathbb{R}[x]$ such that $f(i) = 0$, which one can verify is equal to $(x^2 + 1)$. Therefore, we have $\mathbb{R}[x]/(x^2 + 1) = \mathbb{R}[x]/\mathcal{N}(f) \cong \mathbb{C}$.

Example 15.5.18.  Example 15.4.12 shows that for any $p \in \mathbb{F}$ the kernel of the evaluation map $e_p : \mathbb{F}[x] \rightarrow \mathbb{F}$ is given by $\mathcal{N}(e_p) = (x - p)$. The evaluation map is surjective because for any $a \in \mathbb{F}$ the constant function $a \in \mathbb{F}[x]$ satisfies $e_p(a) = a$. The first isomorphism theorem implies that $\mathbb{F}[x]/(x - p) \cong \mathbb{F}$.

Example 15.5.19.  Example 15.4.11 shows that for any $A \in M_n(\mathbb{C})$ the kernel of the evaluation map $e_A : \mathbb{C}[x] \rightarrow \mathbb{C}[A]$ is given by $\mathcal{N}(e_A) = (p(z))$, where $p(z)$ is the minimal polynomial of A . But the definition of e_A shows that it is surjective onto $\mathbb{C}[A]$, so the first isomorphism theorem gives $\mathbb{C}[x]/(p(z)) \cong \mathbb{C}[A]$.

15.6 The Chinese Remainder Theorem

In this section we discuss the *Chinese remainder theorem* (CRT) and some of its applications. Its earliest version dates back to approximately the fifth century AD and can be found in the book *The Mathematical Classic of Sun Zi*.

The CRT is a powerful tool for working with congruences in the rings $\mathbb{Z}_n$, as well as in polynomial rings, with direct applications in cryptography and coding theory. But it also has many other applications, including partial fraction decompositions and polynomial interpolation. The CRT can be proved for arbitrary (even noncommutative) rings, but all our applications are for Euclidean domains. Since the proofs and results are much cleaner in the case of Euclidean domains, we assume for the rest of the chapter that all rings are Euclidean domains.

The CRT states that given any list of pairwise relatively prime elements $m_1, \dots, m_n$ in the ring R , and given any other elements $a_1, \dots, a_n$, the system of congruences

$$\begin{aligned}
x &\equiv a_1 \pmod{m_1}, \\
x &\equiv a_2 \pmod{m_2}, \\
&\vdots \\
x &\equiv a_n \pmod{m_n}
\end{aligned} \tag{15.3}$$

has a unique solution in $R/(m_1 m_2 \cdots m_n)$. We sometimes call the system (15.3) the *Chinese remainder problem*, or the *CR problem*. It is easy to show that if the solution exists, it is unique. In the case that $R = \mathbb{Z}$, it is also easy to give a nonconstructive proof that there is a solution, but often we want to actually solve the CR problem—not just know that it has a solution. Also, in a more general ring, the nonconstructive proof usually doesn't work.

The way to deal with both issues is to construct a solution to the CR problem. To do this we need either the *Lagrange decomposition* or the *Newton decomposition*. These give an explicit algorithm for solving the CR problem, and as a nice side benefit they also give a proof that every rational function has a partial fraction decomposition.

15.6.1 Lagrange and Newton Decompositions

Before we discuss the CRT, we prove a theorem we call the *Lagrange decomposition* that says how to write elements of a Euclidean domain in terms of pairwise relatively prime elements. This decomposition gives, as a corollary, both the partial fraction decomposition of rational functions and the CRT. We also provide an alternative decomposition due to Newton that can be used for an alternative proof of the CRT and that gives a refined version of the partial fraction decomposition. Both the Lagrange and Newton decompositions can be used to construct polynomial interpolations—polynomials that take on prescribed values at certain points.

Theorem 15.6.1 (Lagrange Decomposition). *Let R be a Euclidean domain, and let $m_1, \dots, m_n$ be pairwise relatively prime elements in R (that is, any m_i and m_j are relatively prime whenever $i \neq j$). Let $M = m_1 m_2 \cdots m_n$. Any element $f \in R$ can be written as*

$$f \equiv \tilde{f}_1 + \cdots + \tilde{f}_n \pmod{M},$$

with each $\tilde{f}_i \equiv 0 \pmod{m_j}$ for $i \neq j$. Alternatively, we can also write

$$f = f_1 + \cdots + f_n + H,$$

with each $f_i \equiv 0 \pmod{m_j}$ and $0 \leq v(f_i) < v(M)$ for every $j \neq i$, and with $H \equiv 0 \pmod{M}$.

Proof of the Lagrange Decomposition Algorithm. For each $i \in \{1, \dots, n\}$ let $\pi_i = \prod_{j \neq i} m_j$. Since all the m_k are pairwise relatively prime, each π_i is relatively prime to m_i , and hence there exist elements $s_i, t_i \in R$ such that $\pi_i s_i + m_i t_i = 1$ (these can be found by the EEA). Let

$$L_i = \pi_i s_i = 1 - m_i t_i.$$

By definition, $L_i \equiv 1 \pmod{m_i}$ and $L_i \equiv 0 \pmod{m_j}$ for every $j \neq i$. Let $L = \sum_{i=1}^n L_i$, so that $L \equiv 1 \pmod{m_i}$ for every $i \in \{1, \dots, n\}$. This means that $1 - L \in (m_i)$ for every i , and by Corollary 15.3.5 the product M must divide $1 - L$.

For all $i \in \{1, \dots, n\}$ let $\tilde{f}_i = f \cdot L_i$, and let $\tilde{H} = f \cdot (1 - L)$, so that we have $f = \sum_{i=1}^n \tilde{f}_i + \tilde{H}$ with $\tilde{f}_i \equiv f \cdot L_i \equiv 0 \pmod{m_j}$ for every $j \neq i$ and $\tilde{H} \equiv 0 \pmod{M}$.

Now using the division property, write each $\tilde{f}_i$ uniquely as $Q_i M + f_i$ for some Q_i and some f_i with $0 \leq v(f_i) < v(m_1 m_2 \cdots m_n)$. This gives

$$f = f_1 + Q_1 M + \cdots + f_n + Q_n M + \tilde{H}.$$

Letting $H = \tilde{H} + M \sum_{i=1}^n Q_i$ gives the desired result. $\square$

Example 15.6.2. To write the Lagrange decomposition of $8 \in \mathbb{Z}$ in terms of $m_1 = 3$, $m_2 = 5$, and $m_3 = 7$, compute $\pi_1 = 5 \cdot 7 = 35$. Using the EEA for 3 and 35, compute $1 = 12 \cdot 3 + (-1) \cdot 35$, so $L_1 = -35$. Similarly, $\pi_2 = 21$ and $1 = (-4) \cdot 5 + 21$, so $L_2 = 21$. Finally $\pi_3 = 15$ and $1 = (-2) \cdot 7 + 15$, so $L_3 = 15$ and

$$1 = -35 + 21 + 15 = L_1 + L_2 + L_3.$$

Multiplication by 8 gives

$$8 = -280 + 168 + 120.$$

Now reduce mod $3 \cdot 5 \cdot 7 = 105$ to get $8 \equiv -70 + 63 + 15 \pmod{105}$. In this special case equality holds: $8 = -70 + 63 + 15$, and we have

$$\begin{aligned} -70 &\equiv 0 \pmod{5} \text{ and } \pmod{7}, \\ 63 &\equiv 0 \pmod{3} \text{ and } \pmod{7}, \\ 15 &\equiv 0 \pmod{5} \text{ and } \pmod{3}. \end{aligned}$$

Remark 15.6.3. The Lagrange decomposition is often used in the special case where $R = \mathbb{F}[x]$ and $m_i = (x - p_i)$ for some collection of distinct points $p_1, \dots, p_n \in \mathbb{F}$. In this case, we do not need the EEA to compute the L_i because there is a simple formula for them:

$$L_i = \prod_{j \neq i} \frac{x - p_j}{p_i - p_j}. \quad (15.4)$$

These are called the *Lagrange interpolants* or *Lagrange basis functions*.

In the proof of the Lagrange decomposition, we have

$$\pi_i = \prod_{j \neq i} m_j = \prod_{j \neq i} (x - p_j),$$

and $\pi_i s_i + m_i t_i = 1$ means precisely that $\pi_i s_i \equiv 1 \pmod{m_i}$; thus, s_i is the multiplicative inverse of π_i after reducing modulo $m_i = (x - p_i)$. As shown in

Example 15.5.18, the first isomorphism theorem says that reducing modulo the ideal $(x - p_i)$ is exactly equivalent to evaluating at the point p_i , thus $\pi_i \equiv \prod_{j \neq i} (p_i - p_j) \pmod{(x - p_i)}$ and $s_i \equiv \prod_{j \neq i} \frac{1}{p_i - p_j}$.

The Newton decomposition gives an alternative way to decompose any element in a Euclidean domain.

Theorem 15.6.4 (Newton Decomposition). *If R is a Euclidean domain with valuation v , and if $m_1, \dots, m_n \in R$ are any elements of R , then for any $f \in R$ we may write*

$$f = b_0 + b_1 m_1 + b_2 m_1 m_2 + \cdots + b_n m_1 m_2 \cdots m_n,$$

with $0 \leq v(b_i) < v(m_{i+1})$ for every $i \in \{0, \dots, n-1\}$.

Proof. Use the division property to write $f = m_1 q_1 + b_0$, with $v(b_0) < v(m_1)$, and then again $q_1 = m_2 q_2 + b_1$, with $v(b_1) < v(m_2)$, and so forth until $q_{n-1} = m_n b_n + b_{n-1}$. This gives

$$\begin{aligned} f &= b_0 + m_1 (b_1 + m_2 (b_2 + m_3 (\cdots + m_{n-1} (b_{n-1} + m_n b_n)))) \\ &= b_0 + b_1 m_1 + b_2 m_1 m_2 + \cdots, \end{aligned}$$

as desired. $\square$

The following is an easy corollary.

Corollary 15.6.5. *For any field $\mathbb{F}$ and any $f, g \in \mathbb{F}[x]$, then for any positive integer n , we may write $f = f_0 + f_1 g + f_2 g^2 + \cdots + f_n g^n$ such that for every $i < n$ we have $\deg(f_i) < \deg(g)$.*

15.6.2 The Chinese Remainder Theorem

As mentioned before, the CR problem is the problem of finding an x such that $x \equiv a_i \pmod{m_i}$ for each i . The Chinese remainder theorem (CRT) guarantees that a solution exists and that it is unique, modulo the product $m_1 m_2 \cdots m_n$. In this section we prove the CRT, but we are interested in more than just the existence of a solution—we want an algorithm for actually constructing a solution. The Lagrange and Newton decompositions both do that.

Theorem 15.6.6 (Chinese Remainder Theorem). *Given pairwise relatively prime elements $m_1, \dots, m_n$ in a Euclidean domain R (that is, if $i \neq j$, then m_i and m_j are relatively prime), let $m = m_1 m_2 \cdots m_n$. The natural ring homomorphism*

$$R/(m) \rightarrow R/(m_1) \times \cdots \times R/(m_n)$$

given by $\llbracket x \rrbracket_{(m)} \mapsto (\llbracket x \rrbracket_{(m_1)}, \dots, \llbracket x \rrbracket_{(m_n)})$ is an isomorphism.

Proof. The natural map $\phi : R \rightarrow R/(m_1) \times \cdots \times R/(m_n)$ given by $x \mapsto (\llbracket x \rrbracket_{(m_1)}, \dots, \llbracket x \rrbracket_{(m_n)})$ is easily seen to be a ring homomorphism. The kernel of ϕ is the set of all x such that $\llbracket x \rrbracket_{(m_j)} = \llbracket 0 \rrbracket_{(m_j)}$ for every $j \in \{1, \dots, n\}$. But $\llbracket x \rrbracket_{(m_j)} = \llbracket 0 \rrbracket_{(m_j)}$

if and only if $x \in (m_j)$, so we have $\mathcal{N}(\phi) = \{x \in R \mid x \in (m_j) \forall j \in \{1, \dots, n\}\} = \bigcap_{j=1}^n (m_j)$. Corollary 15.3.6 gives $\bigcap_{j=1}^n (m_j) = (m)$, and the first isomorphism theorem gives $R/(m) \cong \text{Im}(\phi)$. All that remains is to show that ϕ is surjective—that is, we must show that at least one solution to the CR problem exists.

In the case that $R = \mathbb{Z}$ we can give a nonconstructive proof by simply counting: The image of ϕ is isomorphic to $\mathbb{Z}/(m_1 \cdots m_n)$, so it has exactly $m_1 \cdots m_n$ elements. But it is also a subset of $\mathbb{Z}/(m_1) \times \cdots \times \mathbb{Z}/(m_n)$, which also has $m_1 \cdots m_n$ elements in it. Hence they must be equal, and ϕ must be surjective. But this proof only works when $R/(m_1) \times \cdots \times R/(m_n)$ is finite—it does not work for other important rings like $\mathbb{F}[x]$.

The proof in the general case gives a useful algorithm for solving the CR problem.

Solution of CR Problem by Lagrange Decomposition: Theorem 15.6.1 gives $1 = L_1 + \cdots + L_n + H \in R$, where $L_i \equiv 0 \pmod{m_j}$ for every $j \neq i$, and $H \equiv 0 \pmod{m}$, and hence $L_i \equiv 1 \pmod{m_i}$. Given any element $([a_1]_{(m_1)}, \dots, [a_n]_{(m_n)}) \in R/(m_1) \times \cdots \times R/(m_n)$, let $x = \sum_{i=1}^n a_i L_i \in R$. For each i we have $x \equiv a_i L_i \equiv a_i \pmod{m_i}$. So $\phi(x) = ([a_1]_{(m_1)}, \dots, [a_n]_{(m_n)})$. $\square$

Remark 15.6.7. Here is an equivalent way to state the conclusion of the CRT: Given pairwise relatively prime elements $m_1, \dots, m_n$ in a Euclidean domain R , for any choice $([a_1]_{m_1}, \dots, [a_n]_{m_n}) \in R/(m_1) \times \cdots \times R/(m_n)$, there exists a unique $[x]_m \in R/(m)$ such that $[x]_{m_i} = [a_i]_{m_i}$ for all $i \in \{1, \dots, n\}$. That is, the system of n equations

$$\begin{aligned} x &\equiv a_1 \pmod{m_1}, \\ x &\equiv a_2 \pmod{m_2}, \\ &\vdots \\ x &\equiv a_n \pmod{m_n} \end{aligned} \tag{15.5}$$

has a unique solution for x modulo $m = m_1 m_2 \cdots m_n$.

Remark 15.6.8. Theorem 15.6.6 returns a unique solution modulo m . The solution is not unique in $\mathbb{Z}$ or in R .

Remark 15.6.9 (Solution of CR Problem by Newton Decomposition (Garner's Formula)). Instead of using the Lagrange decomposition to solve the CR problem, we could use the Newton decomposition. Decomposing the unknown x as $x = b_0 + b_1 m_1 + \cdots + b_n m_1 m_2 \cdots m_n$ gives the system

$$\begin{aligned} b_0 &\equiv a_1 \pmod{m_1}, \\ b_0 + b_1 m_1 &\equiv a_2 \pmod{m_2}, \\ &\vdots \\ b_0 + b_1 m_1 + \cdots + b_n m_1 m_2 \cdots m_n &\equiv a_n \pmod{m_n}. \end{aligned} \tag{15.6}$$

Set $b_0 = a_1$. Since m_1 is relatively prime to m_2 , the EEA gives s_1, s_2 such that $s_1 m_1 + s_2 m_2 = 1$, so that $s_1 m_1 \equiv 1 \pmod{m_2}$. Set $b_1 = s_1(a_2 - b_0)$. For each

$i \in \{1, \dots, n-1\}$, since m_{i+1} is relatively prime to $m_1 m_2 \cdots m_i$, we can find s_i such that $s_i m_1 m_2 \cdots m_i \equiv 1 \pmod{m_{i+1}}$. Setting

$$b_i = s_i(a_{i+1} - b_0 - b_1 m_1 - \cdots - b_{i-1} m_1 m_2 \cdots m_{i-1}) \quad (15.7)$$

gives a solution to the CR problem. When $R = \mathbb{Z}$ this method of solving the CR problem is sometimes called *Garner's formula*.

Just as with the Lagrange decomposition, in the case that $R = \mathbb{F}[x]$ and $m_i = x - p_i$ for $p_i \in \mathbb{F}$, we do not need the EEA to find the inverses s_i , since we can write a simple formula for them. Essentially the same argument as in Remark 15.6.3 shows that we can take $s_i = \frac{1}{\prod_{j \leq i} (p_{i+1} - p_j)}$ for each $i \in \{1, \dots, n-1\}$.

Each of these two methods for solving the CR problem has some advantages, depending on the setting and the application.

Example 15.6.10. The following is an ancient problem, dating to sixth century India [Kan88]. “Find the number that if divided by 8 is known to leave 5, that if divided by 9 leaves a remainder 4, and that if divided by 7 leaves a remainder 1.”

Since 7, 8, and 9 are pairwise relatively prime and $7 \cdot 8 \cdot 9 = 504$, the CRT says that for any system of remainders there exists a unique equivalence class $(\text{mod } 504)$ that satisfies the system. We solve the system using first the Lagrange approach and then the Newton–Garner approach.

Lagrange Approach: Let $m_1 = 7$, $m_2 = 8$, and $m_3 = 9$. The Lagrange decomposition algorithm gives $1 \equiv (-3)(72) + (-1)(63) + (-4)(56) \pmod{504}$. The solution of this CR problem is, therefore, $x \equiv 1(-3)(72) + (5)(-1)(63) + (4)(-4)(56) \equiv 85 \pmod{504}$.

Newton–Garner approach: We write $x = b_0 + b_1 m_1 + b_2 m_1 m_2 + b_3 m_1 m_2 m_3$, or $x = b_0 + 7b_1 + 56b_2 + 504b_3$. Since we are only concerned about solutions modulo 504, we need not compute b_3 . From the Garner formula we have $b_0 = a_1 = 1$. We now must find s_1 such that $s_1 m_1 \equiv 1 \pmod{m_2}$. By inspection (or the EEA if necessary) we have $1 = (-1) \cdot 7 + 1 \cdot 8$, so $s_1 = -1$ and $b_1 = -1(a_2 - b_0) = -(5 - 1) = -4$. Again, we must find s_2 such that $s_2 m_1 m_2 \equiv 1 \pmod{m_3}$. By the EEA we have $(-4) \cdot 56 + 25 \cdot 9 = 1$, so $s_2 = -4$ and $b_2 = s_2(a_3 - b_0 - b_1 m_1) = (-4)(4 - 1 - (-4) \cdot 7) = -124$. This gives $x \equiv 1 + (-4) \cdot 7 + (-124) \cdot 56 \pmod{504}$, which again yields $x = 85$ as the smallest positive representative of the equivalence class.

Example 15.6.11. The Lagrange decomposition and the CRT can simplify the process of dividing polynomials, especially when the dividend has no repeated prime factors. For example, consider the problem of computing the remainder of a polynomial f when dividing by $g = (x)(x-1)(x+1)$. If we have already computed the Lagrange interpolants L_i for the prime factors $m_1 = x$, and $m_2 = x-1$ and $m_3 = x+1$ of g , then division by g is fairly quick, using the CRT.

By (15.4) the Lagrange interpolants are

$$L_1 = (x-1)(x+1)/((1)(-1)) = 1 - x^2, \quad (15.8)$$

$$L_2 = (x-0)(x+1)/((1)(2)) = \frac{1}{2}(x^2 + x), \quad (15.9)$$

$$L_3 = (x-1)(x-0)/((-2)(-1)) = \frac{1}{2}(x^2 - x). \quad (15.10)$$

The CRT says that the map

$$\phi : \mathbb{R}[x]/((x)(x-1)(x+1)) \rightarrow \mathbb{R}[x]/(x) \times \mathbb{R}[x]/(x-1) \times \mathbb{R}[x]/(x+1)$$

is an isomorphism, and that the inverse map

$$\psi : \mathbb{R}[x]/(x) \times \mathbb{R}[x]/(x-1) \times \mathbb{R}[x]/(x+1) \rightarrow \mathbb{R}[x]/((x)(x-1)(x+1))$$

is given by $\psi(a, b, c) = aL_1 + bL_2 + cL_3$. This means that $\psi(\phi(f)) \equiv f \pmod{(x)(x-1)(x+1)}$. But $\phi(f) = (e_0(f), e_1(f), e_{-1}(f)) = (f(0), f(1), f(-1))$, so

$$f \equiv f(0)L_1 + f(1)L_2 + f(-1)L_3 \pmod{(x)(x-1)(x+1)}.$$

Since each L_i is already of degree less than g , the sum $f(0)L_1 + f(1)L_2 + f(-1)L_3$ also has degree less than g and hence must be the remainder of $f \pmod{g}$.

So, for example, to compute $f = 3x^5 + 5x^4 + 9x^2 - x + 11 \pmod{g}$, simply compute $f(0) = 11$, $f(1) = 27$, $f(-1) = 23$, and thus $f \equiv 11(1 - x^2) + \frac{27}{2}(x^2 + x) + \frac{23}{2}(x^2 - x) \pmod{g}$.

This method is especially useful when computing several remainders modulo g —finding the Lagrange basis elements is not necessarily much easier than doing a single division, but subsequent divisions can be completed more rapidly.

This method does not work quite so easily if one of the factors m_i is not linear. First, the simple formula (15.4) for the Lagrange basis elements L_i does not apply, so one must use the EEA or Lagrange–Hermite interpolation (see Section 15.7.4) to compute these.

Second, computing the image $\phi(f)$ is also more difficult. For example, if $m_2 = (x-1)^2$, then we could not compute $\phi(f)$ by evaluating f at $x = 1$. Instead we would have to divide f by $(x-1)^2$. But this is still fairly easy, replacing every x in f by $x = (x-1) + 1$ and then expanding. So, to compute $f \pmod{(x-1)^2}$, we write

$$\begin{aligned} f(x) &= x^4 + 5x + 6 = ((x-1) + 1)^4 + 5((x-1) + 1) + 6 \\ &= (x-1)^4 + \binom{4}{1}(x-1)^3 + \binom{4}{2}(x-1)^2 \\ &\quad + \binom{4}{3}(x-1) + 1 + 5(x-1) + 7 \\ &\equiv 9(x-1) + 8 \pmod{(x-1)^2}. \end{aligned}$$

Application 15.6.12 (Radar Systems). The CRT can be used to improve the accuracy of radar systems. When a single radar pulse is sent out and reflected off an object, the distance to that object can be calculated as $ct/2$, where c is the speed of light and t is the time delay between sending the pulse and receiving its reflection. Tracking a moving object requires sending multiple pulses, but then you can't tell which reflection is from which pulse, so you only know the total time t modulo the time Δ between pulses.

One solution to this problem is to send pulses of two or more different frequencies. Since they are different frequencies, you can tell them apart from each other when they reflect. Note that in most cases, electronic systems can only count a finite number of clock cycles per second, so time is essentially measured as an integer number of clock cycles. If there are n different frequencies of pulses, and if Δ_j is the number of clock cycles between pulses of frequency j , then the true time t for one pulse to travel to the object and back is known modulo Δ_j for each $j \in \{1, \dots, n\}$. If $\Delta_1, \dots, \Delta_n$ are pairwise relatively prime, then the CRT gives t modulo $\prod_{j=1}^n \Delta_j$. If this product is known to be much larger than t could possibly be, then the true value of t (and hence the true distance to the object) has been completely determined.

Application 15.6.13 (Parallelizing Arithmetic Computations). Computations on a computer cluster with many processors can be sped up dramatically if the problem can be broken up into many subcomputations that can be done simultaneously (in parallel) on the separate processors and then assembled cheaply into the final answer. The CRT can be used to parallelize arithmetic computations involving very large integers and the ring operations $+$, $-$, $\times$. If primes $p_1, \dots, p_n$ are chosen so that the product $\prod_{i=1}^n p_i$ is certainly larger than the answer to the computation, then the problem can be computed modulo p_i on the i th processor, and then the final answer assembled via the CRT from all these separate pieces.

15.6.3 * Partial Fraction Decomposition

Lagrange and Newton decompositions can be used to prove that every rational function has a partial fraction decomposition.

Corollary 15.6.14 (Partial Fraction Decomposition). *For any field $\mathbb{F}$, if $g_1, \dots, g_n$ are pairwise relatively prime polynomials in $\mathbb{F}[x]$, let $G = \prod_{i=1}^n g_i$. Any rational function of the form f/G with $f \in \mathbb{F}[x]$ can be written as*

$$\frac{f}{G} = h + \frac{s_1}{g_1} + \dots + \frac{s_n}{g_n},$$

with $h, s_1, \dots, s_n \in \mathbb{F}[x]$, and $\deg(s_i) < \deg(g_i)$ for every $i \in \{1, \dots, n\}$.

Proof. The elements g_i satisfy the hypothesis of the Lagrange decomposition theorem (Theorem 15.6.1), so we may write $f = f_1 + \cdots + f_n + H$, with $f_i \equiv 0 \pmod{g_j}$ and $v(f_i) < v(G)$ for all $i \neq j$, and $H \equiv 0 \pmod{G}$.

Thus, we have

$$\frac{f}{G} = \sum_{i=1}^n \frac{f_i}{G} + \frac{H}{G}.$$

The relation $f_i \equiv 0 \pmod{g_j}$ for each $j \neq i$ is equivalent to $g_j | f_i$ for each $j \neq i$. Since the elements g_i are relatively prime, we must have $(\prod_{j \neq i} g_j) | f_i$. Hence $f_i = (\prod_{j \neq i} g_j) s_i$ for some $s_i \in \mathbb{F}[x]$. This gives

$$\frac{f}{G} = \frac{s_1}{g_1} + \cdots + \frac{s_n}{g_n} + \frac{H}{G}.$$

Moreover, we have $G | H$ so $H = hG$, and thus

$$\frac{f}{G} = \frac{s_1}{g_1} + \cdots + \frac{s_n}{g_n} + h.$$

Finally, we have

$$\deg(s_i) = \deg(f_i) - \deg\left(\prod_{j \neq i} g_j\right) < \deg(G) - \deg\left(\prod_{j \neq i} g_j\right) = \deg(g_i). \quad \square$$

The Newton decomposition, or rather its immediate consequence, Corollary 15.6.5, allows further decomposition.

Corollary 15.6.15 (Complete Partial Fraction Decomposition). *Given any polynomials $f, G \in \mathbb{F}[x]$, if G factors as $G = \prod_{i=1}^n g_i^{a_i}$ with all the g_j pairwise relatively prime, then we may write*

$$\frac{f}{G} = h + \left(\frac{s_{11}}{g_1^{a_1}} + \frac{s_{12}}{g_1^{a_1-1}} + \cdots + \frac{s_{1a_1}}{g_1} \right) + \cdots + \left(\frac{s_{n1}}{g_n^{a_n}} + \frac{s_{n2}}{g_n^{a_n-1}} + \cdots + \frac{s_{na_n}}{g_n} \right),$$

with $h \in \mathbb{F}[x]$ and $s_{ij} \in \mathbb{F}[x]$ such that $\deg(s_{ij}) < \deg(g_i)$ for all i and j .

Remark 15.6.16. Constructing the partial fraction decomposition can be done in at least three ways. One way is to explicitly follow the steps in the Lagrange algorithm (the proof of Theorem 15.6.1), using the EEA to write the Lagrange decomposition, which is the main step in the construction of the partial fraction decomposition.

If the denominator G can be written as a square-free product $G = \prod_{i=1}^n (x - p_i)$ with distinct p_i , so each $(x - p_i)$ occurs at most once, then the Lagrange interpolation formula (15.4) gives the Lagrange decomposition of f .

Finally, regardless of the nature of G , one can write the partial fraction decomposition with undetermined coefficients s_{ij} . Clearing denominators and matching the terms of each degree gives a linear system of equations over $\mathbb{F}$, which can be solved for the solution.

15.7 Polynomial Interpolation and Spectral Decomposition

The previous ring-theoretic results lead to some remarkable connections between polynomial interpolation and the spectral decomposition of a matrix. We begin this section with a discussion of interpolation. The connection to spectral decomposition is described in Section 15.7.3.

15.7.1 Lagrange Interpolation

The interpolation problem is this: given points $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{F}^2$, find a polynomial $f(x) \in \mathbb{F}[x]$, such that $f(x_i) = y_i$ for all $i = 1, \dots, n$. As discussed in Example 15.4.12, for any $x_i \in \mathbb{F}$, the kernel of the evaluation homomorphism $e_{x_i} : \mathbb{F}[x] \rightarrow \mathbb{F}$ is the ideal $(x - x_i)$, so we are looking for a polynomial f such that $f \equiv y_i \pmod{(x - x_i)}$ for each $i \in \{1, \dots, n\}$.

The CRT says that there is a unique solution modulo the product $P(x) = \prod_{i=1}^n (x - x_i)$, and one construction of the solution is provided by the Lagrange decomposition. As discussed in Remark 15.6.3, we can use the Lagrange interpolants. This immediately gives the following theorem.

Theorem 15.7.1 (Lagrange Interpolation). *Given points $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{F}^2$ with all the x_i distinct, let*

$$L_i(x) = \prod_{\substack{k=1 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k}. \quad (15.11)$$

The polynomial

$$f(x) = \sum_{i=1}^n y_i L_i(x) \quad (15.12)$$

is the unique polynomial of degree less than n such that $f(x_i) = y_i$ for every $i \in \{1, \dots, n\}$.

Remark 15.7.2. Naïvely computing the polynomials L_i in the form above and using (15.12) is not very efficient—the number of operations required to compute the Lagrange interpolants this way grows quadratically in n . Rewriting the L_i in *barycentric form* makes them much better suited for computation. We write

$$L_i(x) = \frac{p(x)}{(x - x_i)} w_i,$$

where

$$w_i = \prod_{\substack{k=1 \\ k \neq i}}^n \frac{1}{(x_i - x_k)}$$

is the *barycentric weight* and

$$p(x) = \prod_{k=1}^n (x - x_k).$$

This means that

$$f(x) = p(x) \sum_{j=1}^n \frac{w_j}{x - x_j} y_j,$$

which is known as *the first barycentric form*. Moreover, noting that the identity element (the polynomial 1) satisfies

$$1 = p(x) \sum_{j=1}^n \frac{w_j}{x - x_j},$$

we can avoid computing $p(x)$ entirely and write

$$f(x) = \frac{f(x)}{1} = \frac{p(x) \sum_{j=1}^n \frac{w_j}{x - x_j} y_j}{p(x) \sum_{j=1}^n \frac{w_j}{x - x_j}} = \frac{\sum_{j=1}^n \frac{w_j}{x - x_j} y_j}{\sum_{j=1}^n \frac{w_j}{x - x_j}}. \quad (15.13)$$

We call (15.13) the *second barycentric form*. It is not hard to see that the number of computations needed to compute (15.13) grows linearly in n once the weights $w_1, \dots, w_n$ are known. Efficient algorithms exist for adding new points as well; see [BT04] for details.

15.7.2 Newton Interpolation

We can also write a solution to the interpolation problem using the Newton decomposition. Again the proof is an immediate consequence of the CRT and the fact that evaluation of a polynomial at x_j is equivalent to reducing the polynomial modulo $x - x_j$.

Theorem 15.7.3 (Newton Interpolation). *Given points $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{F}^2$, for each $j \in \{0, \dots, n-1\}$ let*

$$\begin{aligned} N_0(x) &= 1, \\ N_j(x) &= \prod_{i \leq j} (x - x_i), \quad \text{and} \\ N_j(x_k) &= \prod_{i \leq j} (x_k - x_i) \end{aligned}$$

and define coefficients $\beta_i \in \mathbb{F}$ recursively, by an analogue of the Garner formula (15.7):

$$\beta_0 = y_1 \quad \text{and} \quad \beta_j = \frac{y_{j+1} - (\beta_0 + \beta_1 n_1(x_{j+1}) + \dots + \beta_{j-1} n_{j-1}(x_{j+1}))}{n_j(x_{j+1})}. \quad (15.14)$$

The polynomial

$$f(x) = \sum_{i=0}^{n-1} \beta_i N_i$$

is the unique polynomial of degree less than n satisfying the conditions $f(x_i) = y_i$ for every $i \in \{1, \dots, n\}$.

Remark 15.7.4. Regardless of whether one uses Newton or Lagrange, the final interpolation polynomial created is the same (provided one uses exact arithmetic). The difference between them is just that they use different bases for the vector space of polynomials of degree less than n . Each of these bases has its own advantages and disadvantages. When computing in floating-point arithmetic the two approaches do not give exactly the same answers, and the Newton approach can have some stability problems, depending on the order of the points $(x_1, y_1), \dots, (x_n, y_n)$. Traditionally, Newton interpolation was preferred in most settings, but more recently it has become clear that barycentric Lagrange interpolation is the better algorithm for many applications (see [BT04]).

Application 15.7.5 (Shamir Secret Sharing). One application of polynomial interpolation and the CRT is the following method for sharing a secret (say, for example, a missile launch code) among a large number of people in such a way that any $k + 1$ of them can deduce the secret, but no smaller number of them can deduce the secret.

To do this choose a random polynomial $f(x)$ of degree k . Tell the first person the value of $f(1)$, the second person the value of $f(2)$, and so forth. Set your secret to be $f(0)$. With only k or fewer people, the value of $f(0)$ cannot be deduced (the missiles cannot be launched), but any $k + 1$ of them will have enough information to reconstruct the polynomial and thus deduce $f(0)$ (and thus launch the missiles).

An alternative method that allows progressively more information to be deduced from each additional collaborator is the following: Choose the secret to be a large integer L and assign each person a prime p_i that is greater than $L^{\frac{1}{k+1}}$ but much smaller than $L^{\frac{1}{k}}$. Let each person know the value of $L \bmod p_i$. If any k of them collaborate (say the first k), they can find the value of $L \bmod \prod_{i=1}^k p_i$, so $L = \ell + m \prod_{i=1}^k p_i$ for some integer $m > 0$, but $\prod_{i=1}^k p_i$ has been chosen to be much smaller than L , so L is not yet uniquely determined. Nevertheless, it is clear that L is now much easier to try to guess than it was before. If $k + 1$ people collaborate, then the value of L is completely known because $\prod_{i=1}^{k+1} p_i > L$.

The two methods of secret sharing are both applications of the CRT, but the first is in the ring $\mathbb{F}[x]$, while the second is in the ring $\mathbb{Z}$.

15.7.3 Lagrange and the Spectral Decomposition

In this section we describe a remarkable connection between the Lagrange interpolants and the spectral decomposition of a matrix.

Consider a matrix $A \in M_n(\mathbb{C})$ with spectrum $\sigma(A)$ and minimal polynomial $p(z) = \prod_{\lambda \in \sigma(A)} (z - \lambda)^{m_\lambda}$, where each m_λ is the order of the eigennilpotent D_λ associated to $\lambda \in \sigma(A)$.

As shown in Examples 15.4.11 and 15.5.19, the map $e_A : \mathbb{C}[x] \rightarrow \mathbb{C}[A] \subset M_n(\mathbb{C})$, given by $f(x) \mapsto f(A)$, is a homomorphism. Also $\mathcal{N}(e_A) = (p(x))$, so the first isomorphism theorem (FIT) guarantees that the induced map $\mathbb{C}[x]/(p(x)) \rightarrow$

$\mathbb{C}[A]$ is an isomorphism. Since each $(z - \lambda)^{m_\lambda}$ is relatively prime to any other $(z - \mu)^{m_\mu}$, we may apply the CRT to see that the natural map $\mathbb{C}[x]/(p(x)) \rightarrow \prod_{\lambda \in \sigma} \mathbb{C}[x]/(x - \lambda)^{m_\lambda}$ is also an isomorphism. Putting these together gives the following diagram:

$$\begin{array}{ccccc}
 \mathbb{C}[x] & \xrightarrow{\quad} & \mathbb{C}[x]/(p(x)) & \xrightarrow[\text{FIT}]{\cong} & \mathbb{C}[A] \subset M_n(\mathbb{C}) \\
 & \searrow \pi & \downarrow \cong \text{CRT} & \swarrow \psi & \\
 & & \prod_{\lambda \in \sigma} \mathbb{C}[x]/(x - \lambda)^{m_\lambda} & &
 \end{array} \tag{15.15}$$

Here the map π takes $f(x)$ to the tuple $(f \bmod (x - \lambda_1)^{m_{\lambda_1}}, \dots, f \bmod (x - \lambda_k)^{m_{\lambda_k}})$, where $\sigma(A) = \{\lambda_1, \dots, \lambda_k\}$. Also, ψ is the map that takes an element $a_0I + a_1A + \dots + a_\ell A^\ell \in \mathbb{C}[A]$ and sends it to $\pi(a_0 + a_1x + \dots + a_\ell x^\ell)$. In the case that A is semisimple ($m_\lambda = 1$ for every $\lambda \in \sigma(A)$), the map π simplifies to $\pi(f) = (f(\lambda_1), \dots, f(\lambda_k))$.

The Lagrange decomposition (Theorem 15.6.1) guarantees the existence of polynomials $L_\lambda \in \mathbb{C}[x]$ for each $\lambda \in \sigma(A)$ such that $\sum_{\lambda \in \sigma(A)} L_\lambda \equiv 1 \bmod p(x)$ and $L_\lambda \equiv 0 \bmod (x - \mu)^{m_\mu}$ for every eigenvalue $\mu \neq \lambda$. This is equivalent to saying that $\pi(L_\lambda) = (0, \dots, 0, 1, 0, \dots, 0)$, where the 1 occurs in the position corresponding to λ . In the case that $m_\lambda = 1$ for every $\lambda \in \sigma(A)$, the L_λ are just given by Theorem 15.7.1,

$$L_\lambda(x) = \prod_{\substack{\mu \in \sigma(A) \\ \mu \neq \lambda}} \frac{x - \mu}{\lambda - \mu},$$

but if A is not semisimple, the multiplicities m_λ are not all 1 and the formula is more complicated (see (15.18) below).

Example 15.7.6. Let

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix},$$

which has minimal polynomial $p(x) = (x - 3)^2(x - 5)$. We have

$$\mathbb{C}[A] \cong \mathbb{C}[x]/((x - 3)^2) \times \mathbb{C}[x]/(x - 5) \cong \{a + b(x - 3) \mid a, b \in \mathbb{C}\} \times \mathbb{C}.$$

The Lagrange interpolants in this case are

$$\begin{aligned}
 L_3 &= -\frac{(x - 3)(x - 5)}{4} - \frac{x - 5}{2} = -(x - 5) \left(\frac{x - 3}{4} + \frac{1}{2} \right) \text{ and} \\
 L_5 &= \frac{(x - 3)^2}{4}.
 \end{aligned}$$

Writing $x - 5 = x - 3 - 2$ shows that

$$\begin{aligned} L_3 &= -\frac{(x-3)(x-3-2)}{4} - \frac{x-3-2}{2} \\ &\equiv \frac{2(x-3)}{4} - \frac{x-3}{2} + 1 \equiv 1 \pmod{(x-3)^2} \text{ and} \\ L_3 &\equiv 0 \pmod{(x-5)}, \end{aligned}$$

while $L_5 \equiv 0 \pmod{(x-3)}$ and $L_5 \equiv 1 \pmod{(x-5)}$. Applying the map e_A to L_3 gives

$$\begin{aligned} e_A(L_3) &= L_3(A) = -(A - 5I)((A - 3I)/4 + I/2) \\ &= \begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left(\frac{1}{4} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix} + \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \right) \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = P_3. \end{aligned}$$

Similarly,

$$e_A(L_5) = \frac{1}{4} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = P_5.$$

Note also that $e_A((x-3)L_3) = (A-3I)P_3 = D_3$ is the eigennilpotent associated to $\lambda = 3$.

The next theorem shows that the appearance of the eigenprojections and eigennilpotents in the previous example is not an accident.

Theorem 15.7.7. *Given the setup described above for a matrix $A \in M_n(\mathbb{C})$, the eigenprojection P_λ is precisely the image of L_λ under the map $e_A : \mathbb{C}[x] \rightarrow \mathbb{C}[A] \subset M_n(\mathbb{C})$. That is, for each $\lambda \in \sigma(A)$ we have*

$$P_\lambda = e_A(L_\lambda) = L_\lambda(A). \quad (15.16)$$

Moreover, the eigennilpotent D_λ is the image of $(x-\lambda)L_\lambda$:

$$D_\lambda = e_A((x-\lambda)L_\lambda) = (A-\lambda I)(L_\lambda(A)). \quad (15.17)$$

Proof. By the uniqueness of the spectral decomposition (Theorem 12.7.5) it suffices to show that the operators $Q_\lambda = e_A(L_\lambda)$ and $C_\lambda = e_A((x-\lambda)L_\lambda)$ satisfy the usual properties:

- (i) $Q_\lambda^2 = Q_\lambda$.
- (ii) $Q_\lambda Q_\mu = 0$ when $\lambda \neq \mu$.
- (iii) $\sum_{\lambda \in \sigma(A)} Q_\lambda = I$.

$$(iv) \quad Q_\lambda C_\lambda = C_\lambda Q_\lambda = C_\lambda.$$

$$(v) \quad A = \sum_{\lambda \in \sigma(A)} (\lambda Q_\lambda + C_\lambda).$$

Since ψ is an isomorphism (see (15.15)), it suffices to verify these properties for $\psi(Q_\lambda) = \pi(L_\lambda)$. For (i) we have

$$\begin{aligned} \pi(L_\lambda)\pi(L_\lambda) &= (0, \dots, 0, 1, 0, \dots, 0)(0, \dots, 0, 1, 0, \dots, 0) \\ &= (0, \dots, 0, 1, 0, \dots, 0) = \pi(L_\lambda), \end{aligned}$$

and a similar argument gives (ii). Item (iii) follows from the fact that $\sum_\lambda L_\lambda \equiv 1 \pmod{p(x)}$ where $p(x)$ is the minimal polynomial of A , and (iv) follows from the fact that

$$\psi(C_\lambda Q_\lambda) = \pi((x - \lambda)L_\lambda)\pi(L_\lambda) = \pi(x - \lambda)\pi(L_\lambda^2) = \pi(x - \lambda)\pi(L_\lambda) = \psi(C_\lambda)$$

and the fact that $\mathbb{C}[A]$ is commutative. Finally, for (v) observe that

$$\begin{aligned} \psi \left(\sum_{\lambda \in \sigma(A)} (\lambda Q_\lambda + C_\lambda) \right) &= \sum_{\lambda \in \sigma(A)} (\pi(\lambda L_\lambda) + \pi((x - \lambda)L_\lambda)) \\ &= \sum_{\lambda \in \sigma(A)} \pi(x L_\lambda) \\ &= \pi(x) \sum_{\lambda \in \sigma(A)} \pi(L_\lambda) \\ &= \pi(x) = \psi(A), \end{aligned}$$

so $\sum_{\lambda \in \sigma(A)} (\lambda Q_\lambda + C_\lambda) = A$. $\square$

Example 15.7.8. Let

$$B = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix},$$

with minimal polynomial $x^2 - 5x$. The CRT gives an isomorphism $\varphi : \mathbb{C}[B] \rightarrow \mathbb{C}[x]/x \times \mathbb{C}[x]/(x-5) \cong \mathbb{C} \times \mathbb{C}$. The Lagrange interpolants are $L_0 = -(x-5)/5$ and $L_5 = x/5$, and again we have $\varphi(L_0) = (1, 0)$ and $\varphi(L_5) = (0, 1)$. Applying the evaluation map gives

$$P_0 = e_B(L_0) = -\frac{1}{5}(B - 5I) = -\frac{1}{5} \begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix}$$

and

$$P_5 = e_B(L_5) = \frac{1}{5}(B).$$

It is easy to see that $P_0^2 = P_0$, that $(\frac{1}{5}B)^2 = \frac{1}{5}B$, and that $BP_0 = 0$, as expected for the eigenprojections.

15.7.4 * Lagrange–Hermite Interpolation

Lagrange interpolation gives the solution of the interpolation problem in the case of distinct values. With just a little more work, these polynomial interpolation results generalize to account for various derivatives at each point as well. This is called *Lagrange–Hermite interpolation*. To begin we need to connect equivalence modulo $(x - p)^m$ to properties of derivatives.

Proposition 15.7.9. *Given $f \in \mathbb{F}[x]$, given $p \in \mathbb{F}$, and given a positive integer n , if $f \equiv 0 \pmod{(x - p)^n}$, then for every nonnegative integer $j < n$, we have*

$$f^{(j)} \equiv 0 \pmod{(x - p)^{n-j}},$$

where $f^{(j)} = \frac{d^j f}{dx^j}$ is the j th derivative of f .

Proof. The statement $f \equiv 0 \pmod{(x - p)^n}$ means that $f = (x - p)^n g$ for some $g \in \mathbb{F}[x]$. Repeated application of the product rule gives

$$f^{(j)} = \sum_{i=0}^j \binom{j}{i} \frac{n!}{(n-i)!} (x - p)^{n-i} g^{(j-i)},$$

from which we immediately see that $(x - p)^{n-j}$ divides $f^{(j)}$, which means that $f^{(j)} \equiv 0 \pmod{(x - p)^{n-j}}$. $\square$

Suppose we want to find a polynomial with certain derivatives at a point. The Taylor polynomial does this at a single point.

Proposition 15.7.10. *Given $f \in \mathbb{F}[x]$ and $p \in F$ and a positive integer n , we have*

$$f \equiv \sum_{i=0}^{n-1} \frac{a_i}{i!} (x - p)^i \pmod{(x - p)^n}$$

if and only if for every nonnegative integer $j < n$ we have $f^{(j)}(p) = a_j$.

Proof. If $f \equiv \sum_{i=0}^{n-1} \frac{a_i}{i!} (x - p)^i \pmod{(x - p)^n}$, then the previous proposition gives

$$\frac{d^j}{dx^j} \left(f - \sum_{i=0}^{n-1} \frac{a_i}{i!} (x - p)^i \right) \equiv 0 \pmod{(x - p)^{n-j}}$$

for every $j < n$. A direct computation gives

$$\frac{d^j}{dx^j} \left(\sum_{i=0}^{n-1} \frac{a_i}{i!} (x - p)^i \right) = \sum_{k=0}^{n-j-1} \frac{a_{j+k}}{k!} (x - p)^k,$$

and hence

$$f^{(j)} - \sum_{k=0}^{n-j-1} \frac{a_{j+k}}{k!} (x-p)^k \equiv 0 \pmod{(x-p)^{n-j}}.$$

Evaluating at $x = p$ gives $f^{(j)}(p) = a_j$.

Conversely, Corollary 15.6.5 guarantees that any polynomial $f \in \mathbb{F}[x]$ can be written as $f = \sum_{i=0}^n b_i(x-p)^i$, where $\deg(b_i) < \deg(x-p)$ for every $i < n$. This shows b_i is constant for each $i < n$. A direct computation of the derivatives of f shows that for each $j < n$ we have $b_j = \frac{a_j}{j!}$; hence $f = (\sum_{i=0}^{n-1} \frac{a_i}{i!} (x-p)^i) + b_n(x-p)^n$ (here $b_n \in \mathbb{F}[x]$ is not necessarily of degree 0). $\square$

Theorem 15.7.11 (Lagrange–Hermite Interpolation). *Given points $x_1, \dots, x_n \in \mathbb{F}$, positive integers $m_1, \dots, m_n$, and values*

$$y_1^{(0)}, \dots, y_1^{(m_1)}, y_2^{(0)}, \dots, y_2^{(m_2)}, \dots, y_n^{(0)}, \dots, y_n^{(m_n)}$$

in $\mathbb{F}$, let $M = \sum_{j=1}^n m_j$, and let $P(x) = \prod_{i=1}^n (x-x_i)^{m_i+1}$. Use the partial fraction decomposition (Corollary 15.6.15) to write

$$\frac{1}{P(x)} = \sum_{i=1}^n \frac{s_{i,0} + s_{i,1}(x-x_i) + \dots + s_{i,m_i}(x-x_i)^{m_i}}{(x-x_i)^{m_i+1}}.$$

For each $i \in \{1, \dots, n\}$ define L_i to be P times the i th term of the partial fraction decomposition,

$$L_i = P(x) \cdot \left(\frac{s_{i,0} + s_{i,1}(x-x_i) + \dots + s_{i,m_i}(x-x_i)^{m_i}}{(x-x_i)^{m_i+1}} \right), \quad (15.18)$$

and define

$$f_i = \sum_{k=0}^{m_i} \frac{y_i^{(k)}}{k!} (x-p)^k.$$

The function

$$f = \sum_{i=1}^n f_i L_i$$

is the unique polynomial in $\mathbb{F}[x]$ of degree less than M such that $f^{(j)}(x_i) = y_i^{(j)}$ for every $i \in \{1, \dots, n\}$ and every $j \in \{0, \dots, m_i\}$.

Proof. By Proposition 15.7.10 the condition that for each $i \leq n$ and each $j \leq m_i$ we have $f^{(j)}(p) = y_i^{(j)}$ is equivalent to the condition that $f \equiv f_i \pmod{(x-x_i)^{m_i+1}}$ for each $i < n$.

By construction, we have $1 = \sum_{i=1}^n L_i$ with $L_i \equiv 0 \pmod{(x-x_j)^{m_j}}$ for any $j \neq i$. Thus, we also have $L_i \equiv 1 \pmod{(x-x_i)^{m_i+1}}$ and $f \equiv f_i \pmod{(x-x_i)^{m_i+1}}$ for any $i \in \{1, \dots, n\}$. Therefore, the derivatives of f take on the required values.

To prove uniqueness, note that the ideals $((x-x_i)^{m_i+1})$ are all pairwise relatively prime and the intersection of these ideals is exactly (P) , hence the CRT guarantees

there is a unique equivalence class $\llbracket g \rrbracket_P$ in $\mathbb{F}[x]/(P)$ such that for every i we have $\llbracket g \rrbracket_{(P)} = \llbracket f_i \rrbracket_{(P)}$. Finally, by Proposition 15.5.5, there is a unique $f \in \mathbb{F}[x]$ with $\deg(f) < \deg(P)$ such that $f \in \llbracket g \rrbracket_{(P)}$. Hence the solution for f is unique. $\square$

Example 15.7.12. Consider the interpolation problem where $x_1 = 1$ and $x_2 = 2$, with $y_1^{(0)} = 1$, $y_1^{(1)} = 2$, and $y_2^{(0)} = 3$. We have $P(x) = (x-1)^2(x-2)$, and a routine calculation shows that the partial fraction decomposition of $1/P$ is

$$\frac{1}{P(x)} = \frac{-x}{(x-1)^2} + \frac{1}{x-2}.$$

Therefore,

$$L_1 = P(x) \cdot \frac{-x}{(x-1)^2} = -x^2 + 2x,$$

$$L_2 = P(x) \cdot \frac{1}{(x-2)} = (x-1)^2.$$

Moreover, we have

$$f_1 = \frac{1}{0!}(x-1)^0 + \frac{2}{1!}(x-1)^1 = 2x - 1,$$

$$f_2 = \frac{3}{0!}(x-2)^0 = 3,$$

and hence

$$f = f_1 L_1 + f_2 L_2 = (2x-1)(-x^2+2x) + 3(x-1)^2 = -2x^3 + 8x^2 + 8x + 3.$$

It is straightforward to check that $f(1) = 1$, $f'(1) = 2$, and $f(2) = 3$, as required.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should *work every exercise* (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with  are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter. The exercises for each section are separated from those for other sections by a horizontal line.

- 15.1. Prove that for any positive integer n the set $\mathbb{Z}_n$ described in Example 15.1.3(iii) satisfies all the axioms of a ring. Prove, also, that multiplication is commutative and that there exists a multiplicative identity element (unity).
- 15.2. Fill in all the details in the proof of Proposition 15.1.9.
- 15.3. Prove that for any ring R , and for any two elements $x, y \in R$ we have $(-x)(-y) = xy$.
- 15.4. In the ring $\mathbb{Z}$, identify which of the following sets is an ideal, and justify your answer:
- (i) The odd integers.
 - (ii) The even integers.
 - (iii) The set $3\mathbb{Z}$ of all multiples of 3.
 - (iv) The set of divisors of 24.
 - (v) The set $\{n \in \mathbb{Z} \mid n = 3k \text{ or } n = 5j\}$ of all multiples of either 3 or 5.
- 15.5. Provide an example showing that the union of two ideals need not be an ideal.
- 15.6. Prove the following:
- (i) The ideal $(3, 5)$ in $\mathbb{Z}$ generated by 3 and 5 is all of $\mathbb{Z}$. Hint: Show $1 \in (3, 5)$.
 - (ii) The ideal $(x^2, x^2 + x, x + 1)$ in $\mathbb{F}[x]$ is all of $\mathbb{F}[x]$.
 - (iii)* In the ring $\mathbb{C}[x, y]$ the ideal $(x^2 - y^3, x - y)$ is a proper subset of the ideal $(x + y, x - 2y)$.
- 15.7.* Prove Proposition 15.1.34.
- 15.8.* Prove Proposition 15.1.35.
- 15.9.* Prove Theorem 15.1.36.
- 15.10.* Prove Proposition 15.1.37.
-
- 15.11. Let $a = 204$ and $b = 323$. Use the extended Euclidean algorithm (EEA) to find $\gcd(a, b)$ in $\mathbb{Z}$ as well as integers s, t such that $as + bt = \gcd(a, b)$. Do the same thing for $a = x^3 - 3x^2 - x + 3$ and $b = x^3 - 3x^2 - 2x + 6$ in $\mathbb{Q}[x]$.
- 15.12. Prove that if p is prime, then every nonzero element $a \in \mathbb{Z}_p$ has a multiplicative inverse. That is, there exists $x \in \mathbb{Z}_p$ such that $ax \equiv 1 \pmod{p}$. Hint: What is $\gcd(a, p)$?
- 15.13. Find the only integer $0 < a < 72$ satisfying $35a \equiv 1 \pmod{72}$. Now find the only integer $0 < b < 72$ satisfying $35b \equiv 67 \pmod{72}$. Hint: Solve $35a + 72x = 1$ with the EEA.
- 15.14. Find a polynomial q of degree 1 or less such that $(x+1)q = x+2 \pmod{x^2+3}$.

- 15.15. Prove that for any composite (nonprime) integer n , the ring $\mathbb{Z}_n$ is not a Euclidean domain. If p is prime, prove that $\mathbb{Z}_p$ is a Euclidean domain. Hint: Use Exercise 15.12.
- 15.16. Prove that $\mathbb{F}[x, y]$ has no zero divisors but is not a Euclidean domain. Hint: What can you say about the ideal (x, y) in $\mathbb{F}[x, y]$?
- 15.17.* Implement the EEA for integers in Python (or your favorite computer language) from scratch (without importing any additional libraries or methods). Your code should accept two integers x and y and return $\gcd(x, y)$, as well as a and b such that $ax + by = \gcd(x, y)$.
-
- 15.18. For p a prime in a Euclidean domain D , prove or disprove each of the following:
- (i) If $a \in D$ and $p|a^k$, then $p^k|a^k$.
 - (ii) If $a, b \in D$ and $p|a^2 + b^2$, then $p|a^2$ and $p|b^2$.
- 15.19. If I is an ideal in a Euclidean domain R , show that there are only finitely many ideals in R that contain I . Hint: $I = (c)$ for some $c \in R$. Consider the divisors of c .
- 15.20. If $a, b \in \mathbb{Z}$ and $3|a^2 + b^2$ prove that $3|a$ and $3|b$. Hint: If 3 does not divide a , then $a = 3k + 1$ or $a = 3k + 2$ for some $k \in \mathbb{Z}$.
- 15.21. If $p \in \mathbb{Z}$ is any positive prime, prove that $\sqrt{p}$ is irrational. Prove that if $p(x)$ is an irreducible polynomial, then $\sqrt{p(x)}$ cannot be written as a rational function in the form $f(x)/g(x)$ with $f, g \in \mathbb{F}[x]$.
- 15.22. Prove Proposition 15.3.8.
- 15.23. Prove that the ring

$$\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\} \subset \mathbb{C}$$

is *not* a Euclidean domain by showing that it does not satisfy the fundamental theorem of arithmetic. Hint: Find two different factorizations of 9 in $\mathbb{Z}[\sqrt{-5}]$.

-
- 15.24. Prove that each of the maps in Example 15.4.3 is a homomorphism.
- 15.25. Prove that if f is surjective, then $\text{Im } f$ is an ideal of S . Give an example of a homomorphism of rings that is not surjective $f : R \rightarrow S$ where this fails.
- 15.26. Let R and S be commutative rings with multiplicative identities 1_R and 1_S , respectively. Let $f : R \rightarrow S$ be a ring homomorphism that is not identically 0.
- (i) Prove that if f is surjective, then $f(1_R) = 1_S$.
 - (ii) Prove that if S has no zero divisors, then $f(1_R) = 1_S$.
 - (iii) Provide an example to show that the previous results are not true if f is not surjective and S has zero divisors.
- 15.27. What is the kernel of each of the following homomorphisms?
- (i) $\mathbb{Z}_{12} \rightarrow \mathbb{Z}_4$ given by $[[x]]_{12} \mapsto [[x]]_4$.
 - (ii) $\mathbb{Z} \hookrightarrow \mathbb{Q}$ is the usual inclusion.
 - (iii) The map $\mathbb{F}[x] \rightarrow C([a, b]; \mathbb{F})$ described in Example 15.4.3(ii).

- 15.28. Prove that if $f : \mathbb{Z} \rightarrow \mathbb{Z}$ is an isomorphism, then f is the identity map ($f(n) = n$ for all $n \in \mathbb{Z}$).
- 15.29. Prove that the map defined in Example 15.4.17(ii) is indeed an isomorphism to $\mathbb{C}$.
- 15.30. Prove Proposition 15.4.20.
- 15.31. Let R and S be rings. What is the kernel of the homomorphism $p : R \times S \rightarrow S$ given by $p(r, s) = s$?

- 15.32. Let $I = (3) \subset \mathbb{Z}_{12}$ be the multiples of 3 in $\mathbb{Z}_{12}$. Write out all the elements of $\mathbb{Z}_{12}/I$ and write out the addition and multiplication tables for $\mathbb{Z}_{12}/I$. Do the same for $I = (8) \subset \mathbb{Z}_{12}$.
- 15.33. Prove that the ring $\mathbb{F}[x]/(x-a)$ is isomorphic to $\mathbb{F}$ for any $a \in \mathbb{F}$.
- 15.34. Prove that the ring $\mathbb{F}[x]/(x-1)^3$ is isomorphic to the set $\{a_0 + a_1(x-1) + a_2(x-1)^2 \mid a_i \in \mathbb{F}\}$ with addition given by

$$\begin{aligned} & (a_0 + a_1(x-1) + a_2(x-1)^2) + (b_0 + b_1(x-1) + b_2(x-1)^2) \\ &= (a_0 + b_0) + (a_1 + b_1)(x-1) + (a_2 + b_2)(x-1)^2 \end{aligned}$$

and with multiplication given by

$$\begin{aligned} & (a_0 + a_1(x-1) + a_2(x-1)^2)(b_0 + b_1(x-1) + b_2(x-1)^2) \\ &= a_0b_0 + (a_1b_0 + a_0b_1)(x-1) + (a_2b_0 + a_1b_1 + a_0b_2)(x-1)^2. \end{aligned}$$

More generally, for any $\lambda \in \mathbb{F}$ show that the ring $\mathbb{F}[x]/((x-\lambda)^n)$ can be written as $\{\sum_{k=0}^{n-1} a_k(x-\lambda)^k \mid a_i \in \mathbb{F}\}$ with the addition given by

$$\left(\sum_{k=0}^{n-1} a_k(x-\lambda)^k \right) + \left(\sum_{k=0}^{n-1} b_k(x-\lambda)^k \right) = \sum_{k=0}^{n-1} (a_k + b_k)(x-\lambda)^k$$

and with multiplication given by

$$\left(\sum_{k=0}^{n-1} a_k(x-\lambda)^k \right) \left(\sum_{k=0}^{n-1} b_k(x-\lambda)^k \right) = \sum_{k=0}^{n-1} \sum_{\substack{j, \ell \\ j+\ell=k}} a_j b_\ell (x-\lambda)^k.$$

- 15.35. If $\lambda \in \mathbb{F}$ and if $\pi : \mathbb{F}[x] \rightarrow \mathbb{F}[x]/((x-\lambda)^n)$ is the canonical epimorphism, then for any $k \in \mathbb{N}$, write $\pi(x^k)$ in the form $\sum_{j=0}^{n-1} a_j(x-\lambda)^j$.
- 15.36. Recall that an idempotent in a ring is an element a such that $a^2 = a$. The element 0 is always idempotent, as is 1, if it exists in the ring. Find at least one more idempotent (not 0 and not 1) in the ring $\mathbb{R}[x]/(x^4 + x^2)$. Also find at least one nonzero nilpotent in this ring.
- 15.37. Prove Lemma 15.5.14.
- 15.38. Prove Theorem 15.5.15.
- 15.39.* Prove that in any commutative ring R the set N of all nilpotents in R forms an ideal of R . Prove that the quotient ring R/N has no nonzero nilpotent elements.

- 15.40.* Prove the *second isomorphism theorem* (compare this to Corollary 2.3.18, which is the vector space analogue of this theorem): Given two ideals I, J of a ring R , the intersection $I \cap J$ is an ideal of the ring I , and J is an ideal of the ring $I + J$. We have

$$(I + J)/J \cong I/(I \cap J).$$

Hint: Show that the obvious homomorphism from I to $(I+J)/J$ is a surjective homomorphism with kernel $I \cap J$.

- 15.41. In each of the following cases, compute the Lagrange and the Newton decompositions for $f \in R$ relative to the elements $m_1, \dots, m_n$:
- (i) $f = 11$ in $R = \mathbb{Z}$, relative to $m_1 = 2, m_2 = 3, m_3 = 5$.
 - (ii) $f = x^4 - 2x + 7$ in $R = \mathbb{R}[x]$, relative to $m_1 = (x - 1), m_2 = (x - 2), m_3 = (x - 3)$. Hint: Consider using the method of Example 15.6.11 to reduce the amount of dividing you have to do.
- 15.42. A gang of 19 thieves has a pile containing fewer than 8,000 coins. They try to divide the pile evenly, but there are 9 coins left over. As a result, a fight breaks out and one of the thieves is killed. They try to divide the pile again, and now they have 8 coins left over. Again they fight, and again one of the thieves dies. Once more they try to divide the pile, but now they have 3 coins left.
- (i) How many coins are in the pile?
 - (ii) If they continue this process of fighting, losing one thief, and redividing, how many thieves will be left when the pile is finally divided evenly with no remainder?
- 15.43. Let $A \in M_n(\mathbb{C})$ be a square matrix with minimal polynomial $p(x)$ and eigenvalues $\lambda_1, \dots, \lambda_k$ with $p(x) = (x - \lambda_1)^{m_1} \cdots (x - \lambda_k)^{m_k}$. Prove that the ring $\mathbb{C}[A]$ is isomorphic to the product of quotient rings $\mathbb{C}[x]/(x - \lambda_1)^{m_1} \times \cdots \times \mathbb{C}[x]/(x - \lambda_k)^{m_k}$. Hint: Use the result of Example 15.5.19.
- 15.44. Fix a positive integer N and let $\omega = e^{-2\pi i/N}$. Prove that the map $\mathcal{F} : \mathbb{C}[t] \rightarrow \mathbb{C} \times \cdots \times \mathbb{C}$, defined by $\mathcal{F}(p(t)) = (p(\omega^0), p(\omega^1), \dots, p(\omega^{N-1}))$, induces an isomorphism of rings from $\mathbb{C}[t]/(t^N - 1)$ to the ring $\mathbb{C}^N = \mathbb{C} \times \cdots \times \mathbb{C}$. This isomorphism is called the *discrete Fourier transform*, and it plays an important role in signal processing and many other applications. Hint: Use the results of Exercise 15.33
- 15.45.* Implement the Lagrange decomposition algorithm for integers in Python (or your favorite computer language) using only your previous implementation of the EEA (Exercise 15.17), without importing any other libraries or methods. Your code should accept an integer x and a tuple $(m_1, \dots, m_n)$ of pairwise-relatively-prime integers and return a tuple $(x_1, \dots, x_n)$ such that $x \equiv \sum_{i=1}^n x_i \pmod{\prod_{i=1}^n m_i}$ and such that $x_i \equiv 0 \pmod{m_j}$ whenever $i \neq j$.
- 15.46.* Implement the Newton–Garner algorithm for solving the CR problem for integers in Python (or your favorite computer language) using only your previous implementation of the EEA (Exercise 15.17), without importing

any other libraries or methods. Your code should accept a tuple $(a_1, \dots, a_n)$ of integers and a tuple $(m_1, \dots, m_n)$ of pairwise-relatively-prime integers and return an integer $0 \leq x < \prod_{i=1}^n |m_i|$ such that $x \equiv a_i \pmod{m_i}$ for every i .

15.47. Find a polynomial $f(x) \in \mathbb{Q}[x]$ such that $f(1) = 2$, $f(2) = 3$, $f(3) = 5$, and $f(4) = 7$ using

- (i) Lagrange interpolation and
- (ii) Newton interpolation.

15.48. Let

$$A = \begin{bmatrix} 7 & 1 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

- (i) Find the minimal polynomial of A .
- (ii) Write a ring of the form $\prod_{\lambda \in \sigma(A)} \mathbb{C}[x]/(x - \lambda)^{m_\lambda}$ that is isomorphic to $\mathbb{C}[A]$.
- (iii) Compute the Lagrange–Hermite interpolants L_λ for each $\lambda \in \sigma(A)$. Verify that $L_\lambda \equiv 0 \pmod{(x - \mu)^{m_\mu}}$ for every $\mu \neq \lambda$, and that $L_\lambda \equiv 1 \pmod{(x - \lambda)^{m_\lambda}}$.
- (iv) Compute the eigenprojections $P_\lambda = e_A(L_\lambda) \in M_n(\mathbb{C})$ by evaluating the Lagrange–Hermite interpolants at A .
- (v) Compute the eigennilpotents D_λ in a similar fashion; see (15.17).

15.49. Let

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}.$$

- (i) Find the minimal polynomial of B . Hint: Do not compute the resolvent.
 - (ii) Write a ring of the form $\prod_{\lambda \in \sigma(A)} \mathbb{C}[x]/(x - \lambda)^{m_\lambda}$ that is isomorphic to $\mathbb{C}[B]$.
 - (iii) Compute the Lagrange–Hermite interpolants L_λ for each $\lambda \in \sigma(B)$. Verify that $L_\lambda \equiv 0 \pmod{(x - \mu)^{m_\mu}}$ for every $\mu \neq \lambda$, and that $L_\lambda \equiv 1 \pmod{(x - \lambda)^{m_\lambda}}$.
 - (iv) Compute the eigenprojections $P_\lambda = e_B(L_\lambda) \in M_n(\mathbb{C})$ by evaluating the Lagrange–Hermite interpolants at B .
 - (v) Compute the eigennilpotents D_λ in a similar fashion; see (15.17).
- 15.50. Given $A \in M_n(\mathbb{C})$, with spectrum $\sigma(A)$ and minimal polynomial $p(x) = \prod_{\lambda \in \sigma(A)} (x - \lambda)^{m_\lambda}$, for each $\lambda \in \sigma(A)$, let L_λ be the corresponding polynomials in the Lagrange decomposition with $L_\lambda \equiv 0 \pmod{(x - \mu)^{m_\mu}}$ for every eigenvalue $\mu \neq \lambda$, and with $\sum_{\lambda \in \sigma(A)} L_\lambda \equiv 1 \pmod{p(x)}$. Use the results of this section to show that the Drazin inverse A^D lies in $\mathbb{C}[A] \subset M_n(\mathbb{C})$, and that

$$A^D = e_A \left(\sum_{\lambda \in \sigma(A)} L_\lambda \sum_{k=0}^{m_\lambda-1} \frac{(x - \lambda)^k (-1)^k}{\lambda^{k+1}} \right),$$

where $e_A : \mathbb{C}[x] \rightarrow \mathbb{C}[A]$ is the evaluation map given by $x \mapsto A$.

- 15.51. Use the techniques of this section (polynomial interpolation) to give a formula for the inverse of the discrete Fourier transform of Exercise 15.44.

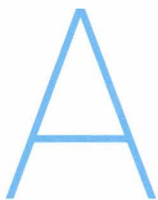
Notes

Some references for the material in this chapter include [Art91, Her96, Alu09]. Several of the applications in this chapter are from [Wik14].

Variations of Exercise 15.42 date back to Qin Jiushao's book *Shushu Jiuzhang* (Nine Sections of Mathematics), written in 1247 (see [Lib73]), but we first learned of this problem from [Ste09].

Part V

Appendices



Foundations of Abstract Mathematics

It has long been an axiom of mine that the little things are infinitely the most important.

—Sir Arthur Conan Doyle

In this appendix we provide a quick sketch of some of the standard foundational ideas of mathematics, including the basics of set theory, relations, functions, orderings, induction, and Zorn’s lemma. This is not intended to be a complete treatment of these topics from first principles—for that see [GC08, DS10, Hal74]. Our purpose here is to provide a handy reference for some of these basic ideas, as well as to solidify notation and conventions used throughout the book.

A.1 Sets and Relations

A.1.1 Sets

Definition A.1.1. *A set is an unordered collection of distinct elements, usually indicated with braces, as in $S = \{A, 2, \text{Cougar}, \sqrt{3}\}$. We write $x \in S$ to indicate that x is an element of the set S , and $y \notin S$ to indicate that y is not an element of S . Two sets are considered to be equal if each has precisely the same elements as the other.*

Example A.1.2. Here are some sets that we use often in this book:

- (i) The set with no elements is the *empty set* and is denoted $\emptyset$. The empty set is unique; that is, any set with no elements must be equal to $\emptyset$.
- (ii) The set $\mathbb{C}$ of complex numbers.
- (iii) The set $\mathbb{R}$ of real numbers.

- (iv) The set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ of natural numbers.
- (v) The set $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ of integers.
- (vi) The set $\mathbb{Z}^+ = \{1, 2, 3, \dots\}$ of positive integers.
- (vii) The set $\mathbb{Q}$ of rational numbers.
- (viii) The set $\mathbb{R}^2$ of all pairs (a, b) , where a and b are any elements of $\mathbb{R}$.

Example A.1.3.

- (i) Sets may have elements which are themselves sets. The set

$$A = \{\{1, 2, 3\}, \{r, s, t, u\}\}$$

has two elements, each of which is itself a set.

- (ii) The empty set may be an element of another set. The set $T = \{\emptyset, \{\emptyset\}\}$ has two elements, the first is the empty set and the second is a set $B = \{\emptyset\}$ whose only element is the empty set. If this is confusing, it may be helpful to think of the empty set as a bag with nothing in it, and the set T as a bag containing two items: one empty bag and one other bag B with an empty bag inside of B .

Definition A.1.4. We say that a set T is a subset of a set S , and we write $T \subset S$ if all of the elements of T are also elements of S . We say that T is a proper subset of S if $T \subset S$, and $T \neq \emptyset$, and there is at least one element of S that is not in T .

Example A.1.5.

- (i) The integers are a subset of the rational numbers, which are a subset of the real numbers, which are a subset of the complex numbers:

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

- (ii) The empty set $\emptyset$ is a subset of every set.
- (iii) Given a set S , the *power set* of S is the set of all subsets of S . It is sometimes denoted $\mathcal{P}(S)$ or 2^S . For example, $S = \{a, b, c\}$; then the power set is

$$\mathcal{P}(\{a, b, c\}) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Note that if S is a finite set with n elements, then the power set consists of 2^n elements.

The fact that two sets are equal if and only if they have the same elements leads to the following elementary, but useful, way to prove that two sets are equal.

Proposition A.1.6. *Sets A and B are equal if and only if both $A \subset B$ and $B \subset A$.*

Definition A.1.7. *We often use the following convenient shorthand for writing a set in terms of the properties its elements must satisfy. If P is some property or formula with a free variable x , then we write*

$$\{x \in S \mid P(x)\}$$

to indicate the subset of S consisting of all elements in S which satisfy P . We call this set builder notation or set comprehension. When P contains symbols that look like $\mid$ we use the alternative form $\{x \in S : P(x)\}$ to avoid confusion.

Example A.1.8.

- (i) $\{x \in \mathbb{Z} \mid x > 5\}$ is the set of all integers greater than 5.
- (ii) $\{x \in \mathbb{R} \mid x^2 = 1\}$ is the set of all real numbers whose square is 1.
- (iii) $\{(a, b) \in \mathbb{R}^2 \mid a^3 = b^5\}$ is the set of all pairs $(a, b) \in \mathbb{R}^2$ such that $a^3 = b^5$.
- (iv) $\{x \in \mathbb{R} \mid x^2 = -1\} = \emptyset$ is empty because there are no choices of x that satisfy the condition $x^2 = -1$.

Remark A.1.9. In many cases it is useful to use set comprehensions that do not specify the superset from which the elements are selected, for example, $\{(r, s, t) \mid f(r, s, t) = 0\}$. This can be very handy in some situations, but does have some serious potential pitfalls. First, there is a possibility for misunderstanding. But even when the meaning seems completely clear, this notation can lead to logical paradoxes. The most famous of these is Russell's paradox, which concerns the comprehension $R = \{A \mid A \text{ is a set, and } A \notin A\}$. If R itself is a set, then we have a paradox in the question of whether R contains itself or not.

A proper treatment of these issues is beyond the scope of this appendix and does not arise in most applications. The interested reader is encouraged to consult one of the many standard references on set theory and logic, for example, [Hal74].

Definition A.1.10. *There are several standard operations on sets for building new sets from old.*

- (i) *The union of two sets A and B is*

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

- (ii) *If $\mathcal{A}$ is a set of sets, the union of all the sets in $\mathcal{A}$ is*

$$\bigcup_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

(iii) The intersection of two sets A and B is

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

(iv) If $\mathcal{A}$ is a set of sets, the intersection of all the sets in $\mathcal{A}$ is

$$\bigcap_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for every } A \in \mathcal{A}\}.$$

(v) The set difference $A \setminus B$ is

$$A \setminus B = \{x \in A \mid x \notin B\}.$$

(vi) The complement of a subset $A \subset S$ is

$$A^c = S \setminus A.$$

Note that writing A^c only makes sense if the superset S is already given.

(vii) The Cartesian product (or simply the product) $A \times B$ of two sets A and B is the set of all ordered pairs (a, b) , where $a \in A$ and $b \in B$:

$$A \times B = \{(a, b) \mid a \in A, b \in B\}.$$

(viii) If $\mathcal{A} = [A_1, \dots, A_n]$ is a finite (ordered) list of sets, the product of all the sets in $\mathcal{A}$ is

$$\prod_{A_i \in \mathcal{A}} A_i = \prod_{i=1}^n A_i = \{(x_1, \dots, x_n) \mid x_i \in A_i \text{ for each } i \in \{1, 2, \dots, n\}\}.$$

The product $\underbrace{S \times \dots \times S}_n$ of a set S with itself n times is often written S^n .

Proposition A.1.11 (De Morgan's Laws). If $A, B \subset S$, then

$$(A \cup B)^c = A^c \cap B^c \tag{A.1}$$

and

$$(A \cap B)^c = A^c \cup B^c.$$

More generally, if $\mathcal{A}$ is a set of sets, then

$$\left(\bigcup_{A \in \mathcal{A}} A \right)^c = \left(\bigcap_{A \in \mathcal{A}} A^c \right)$$

and

$$\left(\bigcap_{A \in \mathcal{A}} A \right)^c = \left(\bigcup_{A \in \mathcal{A}} A^c \right).$$

Proof. We will prove (A.1). The proofs of the rest of the laws are similar. By definition, $x \in (A \cup B)^c$ if and only if $x \in S$ and $x \notin A \cup B$, which holds if and only if $x \notin A$ and $x \notin B$. But this is the definition of $x \in A^c \cap B^c$. $\square$

A.1.2 Relations, Equivalence Relations, and Partitions

Definition A.1.12. A relation on a set A is a subset $R \subset A \times A$. We say “ a is related to b ” when we mean $(a, b) \in R$. We usually choose some symbol (for example $\star$) to denote the relation and write $a \star b$ to denote $(a, b) \in R$.

Example A.1.13.

- (i) The less-than symbol defines a relation on $\mathbb{R}$, namely, the subset $L = \{(x, y) \in \mathbb{R}^2 \mid x < y\} \subset \mathbb{R}^2$. So we have $x < y$ if $(x, y) \in L$.
- (ii) Define a relation on $\mathbb{Z}$ by the subset

$$\{(a, b) \mid a \text{ divides } b\} \subset \mathbb{Z} \times \mathbb{Z}.$$

It is normal to use the symbol $\mid$ to denote this relation, so $a \mid b$ if a divides b .

- (iii) Given any set A , define a relation by $E = \{(x, x) \mid x \in A\} \subset A \times A$. This matches the usual definition of equality in A . That is, $(a, b) \in E$ if $a = b$.
- (iv) Fix an integer n and define a relation on $\mathbb{Z}$ by $a \equiv b$ if n divides $(a - b)$. It is traditional to denote this equivalence relation by $a \equiv b \pmod{n}$ and to say that a is congruent (or equivalent) to b modulo n .
- (v) Let F be the set of formal symbols of the form a/b , where $a, b \in \mathbb{Z}$ and $b \neq 0$. More precisely, we have $F = \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$. Define a relation $\sim$ on F by $a/b \sim c/d$ if $ad = bc$. This is the usual relation for equality of fractions.

As we saw in Example A.1.13(iii), equality defines a relation on any set. In many situations we’d like to say that certain elements of a set are alike in some way, but without necessarily saying that they are equal. The definition of “alike” also defines a relation on the set. The idea of *equivalence relation* identifies some key properties that make such relations behave like equality.

Definition A.1.14. Given a relation $R \subset A \times A$, denoted by $\sim$, we say that R (or $\sim$) is an equivalence relation if the following three properties hold:

- (i) (Reflexivity.) For every $a \in A$ we have $a \sim a$.
- (ii) (Symmetry.) Whenever $a \sim b$, then $b \sim a$.
- (iii) (Transitivity.) If $a \sim b$ and $b \sim c$ then $a \sim c$.

Unexample A.1.15.

- (i) The relation $<$ of Example A.1.13(i) is not an equivalence relation on $\mathbb{R}$ because it fails to be either reflexive or symmetric. It is, however, transitive.
- (ii) The relation $|$ of Example A.1.13(ii) is reflexive and (vacuously) transitive. But it is not an equivalence relation because it fails to be symmetric.

Example A.1.16.

- (i) The relation $=$ of Example A.1.13(iii) is an equivalence relation. This is the example that motivated the definition of equivalence relation.
- (ii) The relation $\equiv$ of Example A.1.13(iv) is an equivalence relation. It is easy to see that it is reflexive and symmetric. To see that it is transitive, note that if $a \equiv b \pmod{n}$, then $n|(b-a)$ and $b = kn + a$ for some integer k . If also $b \equiv c \pmod{n}$, then $n|(c-b)$ and $c = n\ell + b$ for some integer ℓ . Thus we have $c = n\ell + b = n\ell + nk + a$, so $c - a = n(\ell + k)$ and $a \equiv c$.
- (iii) The relation on $\mathbb{C}$ given by $x \sim y$ if $|x| = |y|$ is an equivalence relation.
- (iv) The relation of Example A.1.13(v) is an equivalence relation.

It is often very useful to think about equivalence relations in terms of the subsets of equivalent elements it defines.

Definition A.1.17. *Given an equivalence relation $\sim$ on a set S , and given an element $x \in S$, the equivalence class of x is the set $\llbracket x \rrbracket = \{y \in S \mid y \sim x\}$.*

Example A.1.18.

- (i) Fix any integer n , and let $\equiv$ be the equivalence relation $\equiv \pmod{n}$ of Example A.1.13(iv). The equivalence classes are

$$\begin{aligned}
 \llbracket 0 \rrbracket &= \{\dots, -2n, -n, 0, n, 2n, 3n, \dots\}, \\
 \llbracket 1 \rrbracket &= \{\dots, 1 - 2n, 1 - n, 1, 1 + n, 1 + 2n, 1 + 3n, \dots\}, \\
 \llbracket 2 \rrbracket &= \{\dots, 2 - 2n, 2 - n, 2, 2 + n, 2 + 2n, 2 + 3n, \dots\}, \\
 &\vdots \\
 \llbracket n-1 \rrbracket &= \{\dots, -n-1, -1, n-1, 2n-1, 3n-1, \dots\}, \\
 \llbracket n \rrbracket &= \{\dots, -2n, -n, 0, n, 2n, 3n, \dots\} = \llbracket 0 \rrbracket.
 \end{aligned}$$

That is, for each $i \in \mathbb{Z}$, we have $\llbracket i \rrbracket = \{x \in \mathbb{Z} : n \mid (x - i)\}$, and there are precisely n distinct equivalence classes. We denote this set of n equivalence classes by $\mathbb{Z}_n$.

As a special case, if we take $n = 2$, then the two equivalence classes of $\mathbb{Z}_2$ are the even integers and the odd integers.

- (ii) If the equivalence relation on S is equality ($=$), then the equivalence classes are just the singleton sets $\llbracket x \rrbracket = \{x\}$.
- (iii) Consider the equivalence relation on $\mathbb{C}$ given by $x \sim y$ if $|x| = |y|$. For each nonnegative real number r there is exactly one equivalence class $\llbracket r \rrbracket$, consisting of the circle $\llbracket r \rrbracket = \{z \in \mathbb{C} : |z| = r\}$. Every complex number z lies in one of the equivalence classes $\llbracket r \rrbracket$, and no two distinct nonnegative real numbers lie in the same equivalence class.
- (iv) For the relation of Example A.1.13(v) the equivalence class of a given element a/b consists of sets of all equivalent fractions. For example, we have $\llbracket 1/2 \rrbracket = \{1/2, 2/4, 3/6, -1/-2, 57/114, \dots\}$.

The properties of an equivalence relation immediately give the following properties of equivalence classes.

Proposition A.1.19. *Given an equivalence relation $\sim$ on a set S , we have the following:*

- (i) *Any $x \in S$ is contained in the equivalence class $\llbracket x \rrbracket$.*
- (ii) *If $y \in \llbracket x \rrbracket$, then $x \in \llbracket y \rrbracket$.*
- (iii) *Any two equivalence classes are either disjoint or equal; that is, if $y \in \llbracket x \rrbracket \cap \llbracket z \rrbracket$, then $\llbracket x \rrbracket = \llbracket z \rrbracket$.*

Proof. Items (i) and (ii) follow immediately from the definition of an equivalence relation. To prove item (iii), note that for any $a \in \llbracket x \rrbracket$ we have $a \sim x$ and $y \sim x$, so by symmetry and transitivity $a \sim y$. But because $y \in \llbracket z \rrbracket$, we also have $y \sim z$ and hence $a \sim z$, which shows that $a \in \llbracket z \rrbracket$. This shows that $\llbracket x \rrbracket \subset \llbracket z \rrbracket$. An essentially identical argument shows that $\llbracket z \rrbracket \subset \llbracket x \rrbracket$, and hence that $\llbracket x \rrbracket = \llbracket z \rrbracket$. $\square$

Equivalence classes divide a set into disjoint subsets, which we call a partition.

Definition A.1.20. *A partition of a set S is a set $\mathcal{A}$ of nonempty disjoint subsets of S such that*

- (i) $S = \bigcup_{A \in \mathcal{A}} A$ and
- (ii) *for every $A, B \in \mathcal{A}$, if $A \neq B$, then $A \cap B = \emptyset$.*

Example A.1.21.

- (i) The integers have a partition into the even and odd integers. The partition is $\mathcal{A} = \{E, O\}$ with $E = \{\dots, -4, -2, 0, 2, 4, \dots\}$ and $O = \{\dots, -3, -1, 1, 3, 5, \dots\}$. This is a partition because $\mathbb{Z} = E \cup O$ and $E \cap O = \emptyset$.
- (ii) Taking $\mathcal{A} = \{S\}$ gives a (not very interesting) partition of S into just one set—itsself.
- (iii) If $\sim$ is any equivalence relation on a set S , then the equivalence classes of $\sim$ define a partition of S by $\mathcal{A} = \{[x] \mid x \in S\}$. The first condition, that $\bigcup_{A \in \mathcal{A}} A = S$, follows immediately from Proposition A.1.19(i). The fact that the equivalence classes are disjoint is Proposition A.1.19(iii).
- (iv) Taking $\mathcal{A} = \{\{x\} \mid x \in S\}$ gives a partition of S into singleton sets. This is the partition induced by the equivalence relation $=$ of Example A.1.13(iii).
- (v) For any fixed integer n , the collection $\mathcal{A} = \mathbb{Z}_n = \{[0], [1], \dots, [n-1]\}$, with $[i] = \{x \in \mathbb{Z} : n \mid (x - i)\}$, is a partition of $\mathbb{Z}$ because $\bigcup_{i=0}^{n-1} [i] = \mathbb{Z}$ and $[i] \cap [j] = \emptyset$ for every $i \neq j$. This is exactly the partition into equivalence classes (mod n) described in Example A.1.18(i).

Any partition defines an equivalence relation, just as any equivalence relation determines a partition into equivalence classes. It is not hard to see that partitions and equivalence relations are equivalent; that is, each is completely determined by the other.

Proposition A.1.22. *A partition $\mathcal{A}$ of a set S defines an equivalence relation on S by $x \sim y$ if and only if $x, y \in A$ for some $A \in \mathcal{A}$. That is, $x \sim y$ if and only if they both lie in the same part. Moreover, a partition is completely determined by the equivalence relation it defines, and an equivalence relation is completely determined by the corresponding partition into equivalence classes.*

Proof. Given a partition $\mathcal{A}$ of S , let $\sim$ be the induced relation. It is immediate that $\sim$ is both reflexive and symmetric. If $x \sim y$ and $y \sim z$, then there is a part $A \in \mathcal{A}$ such that $x, y \in A$, and a part $B \in \mathcal{A}$ such that $y, z \in B$, but since partitions are disjoint, the fact that $y \in A$ and $y \in B$ implies that $A = B$, so $x, y, z \in A = B$ and $x \sim z$. Thus $\sim$ is an equivalence relation.

Given any two partitions $\mathcal{A} \neq \mathcal{B}$, there must be at least one nonempty set $A \in \mathcal{A} \setminus \mathcal{B}$. Let $x \in A$. Since the partition $\mathcal{B}$ covers all of S , we must have some $B \in \mathcal{B}$ such that $x \in B$. Moreover, since $A \notin \mathcal{B}$, we must have $A \neq B$; and so there must be some $y \in A \setminus B$ or there must be a $z \in B \setminus A$. Without loss of generality, assume $y \in A \setminus B$. Then we have $x \sim_{\mathcal{A}} y$ but $x \not\sim_{\mathcal{B}} y$, so the induced equivalence relations are not the same.

Conversely, given any two equivalence relations $\sim$ and $\equiv$, if they are not the same, there must be a pair of elements x, y such that $x \sim y$, but $x \not\equiv y$. This implies

that y is an element of the $\sim$ equivalence class $[[x]]_{\sim}$ of x , but y is not in the $\equiv$ equivalence class $[[x]]_{\equiv}$ of x . Thus the corresponding partitions are not equal. $\square$

A.2 Functions

Functions are a fundamental notion in mathematics. The formal definition of a function is given in terms of what is often called the graph of the function.

A.2.1 Basic Properties of Functions

Definition A.2.1. A function (also called a map or a mapping) $f : X \rightarrow Y$ is a relation $\Gamma_f \subset X \times Y$ such that for every $x \in X$ there is a unique $y \in Y$ such that $(x, y) \in \Gamma_f$. For each x , we denote this uniquely determined y by $f(x)$. It is traditional to call the set Γ_f the graph of f , although formally speaking, the graph is the function.

We also have the following sets determined by the function f :

- (i) The set X is the domain of f .
- (ii) The set Y is the codomain of f .
- (iii) The set $f(X) = \{y \in Y \mid \exists x \in X \text{ s.t. } f(x) = y\}$ is the range or image of f .
- (iv) For any subset $X' \subset X$, we call the set $f(X') = \{y \in Y \mid \exists x' \in X' \text{ s.t. } y = f(x')\}$ the image of X' .
- (v) For any $y \in Y$, the set $f^{-1}(y) = \{x \in X \mid f(x) = y\}$ is the level set of f at y or the preimage of y .
- (vi) For any $Y' \subset Y$, we call the set $f^{-1}(Y') = \{x \in X \mid f(x) \in Y'\}$ the preimage of Y' .
- (vii) If $0 \in Y$ (for example, if $Y = \mathbb{R}$ or $Y = \mathbb{C}^n$), then the set $\{x \in X \mid f(x) \neq 0\}$ is called the support of f .

Notation A.2.2. As an alternative to the notation $f(x) = y$, it is common to write $f : x \mapsto y$.

The following proposition is immediate from the definition but still very useful.

Proposition A.2.3. Two functions f, g are equal if and only if they have the same domain D and for each $x \in D$ we have $f(x) = g(x)$.

The following proposition is also immediate from the definitions.

Proposition A.2.4. Let $f : X \rightarrow Y$ be a function. For any $R, S \subset X$ and $T, U \subset Y$, the following hold:

- (i) If $R \subset S$, then $f(R) \subset f(S)$.
- (ii) $f(R \cup S) = f(R) \cup f(S)$.

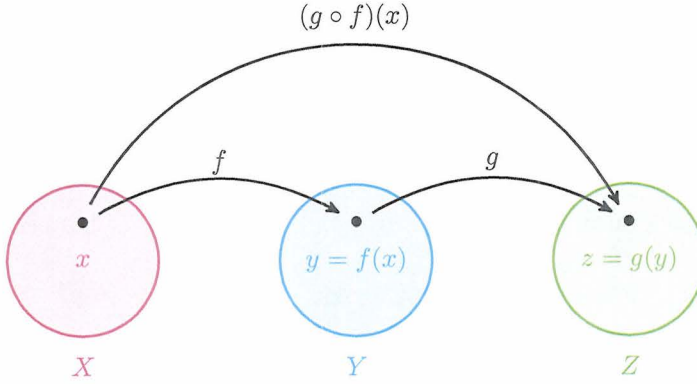


Figure A.1. The composition of $g \circ f$ of functions f and g , as described in Definition A.2.5.

- (iii) $f(R \cap S) \subset f(R) \cap f(S)$.
- (iv) $R \subset f^{-1}(f(R))$.
- (v) If $T \subset U$, then $f^{-1}(T) \subset f^{-1}(U)$.
- (vi) $f^{-1}(T \cup U) = f^{-1}(T) \cup f^{-1}(U)$.
- (vii) $f^{-1}(T \cap U) = f^{-1}(T) \cap f^{-1}(U)$.
- (viii) $f(f^{-1}(T)) \subset T$.
- (ix) $f^{-1}(T^c) = f^{-1}(T)^c$.

Definition A.2.5. The composition of two functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ is the function $g \circ f: X \rightarrow Z$ given by

$$\Gamma_{g \circ f} = \{ (x, z) \mid \exists y \in Y \text{ s.t. } (x, y) \in \Gamma_f \text{ and } (y, z) \in \Gamma_g \}.$$

This is illustrated in Figure A.1.

Proposition A.2.6. The composition $g \circ f$ of functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ is a function $g \circ f: X \rightarrow Z$. Also, composition of functions is associative, meaning that given $f: X \rightarrow Y$, $g: Y \rightarrow Z$ and $h: Z \rightarrow W$, we have

$$h \circ (g \circ f) = (h \circ g) \circ f$$

Proof. Because f is a function, given any $x \in X$, there is a unique $y \in Y$ such that $(x, y) \in \Gamma_f$. Since g is a function, there is a unique $z \in Z$ such that $(y, z) \in \Gamma_g$; hence there is a unique $z \in Z$ such that $(x, z) \in \Gamma_{g \circ f}$.

For the second part of the proposition, note that by Proposition A.2.3 it suffices to show that for any $x \in X$, we have $(h \circ (g \circ f))(x) = ((h \circ g) \circ f)(x)$. But by definition, we have

$$(h \circ (g \circ f))(x) = h((g \circ f)(x)) = h(g(f(x))) = (h \circ g)(f(x)) = ((h \circ g) \circ f)(x). \quad \square$$

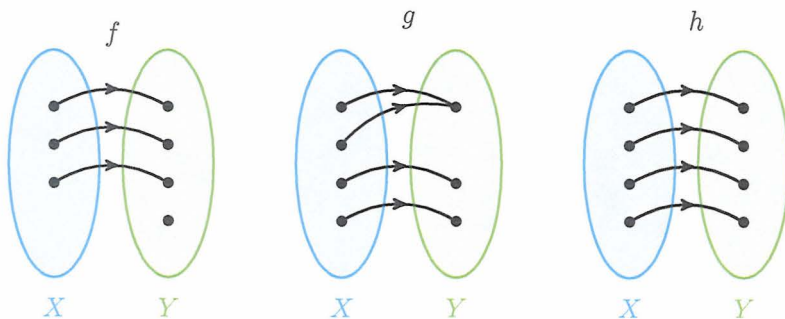


Figure A.2. *Injective, surjective, and bijective maps, as described in Definition A.2.7. The left-hand map f is injective but not surjective. The center map g is surjective but not injective. The right-hand map h is bijective.*

A.2.2 Injective and Surjective

Definition A.2.7. A function $f : X \rightarrow Y$ is called

- (i) injective if $f(x) = f(x')$ implies $x = x'$;
- (ii) surjective if for any $y \in Y$ there exists an $x \in X$ such that $f(x) = y$;
- (iii) bijective if it is both injective and surjective.

See Figure A.2 for an illustration of these ideas.

Nota Bene A.2.8. Many people use the phrase *one-to-one* to mean *injective* and *onto* to mean *surjective*. But the phrase *one-to-one* is misleading and tends to make students mistakenly think of the uniqueness condition for a function rather than the correct meaning of injective. Moreover, most students who have encountered the phrase *one-to-one* before have heard various intuitive definitions for this phrase that are too sloppy to actually use for proofs and that tend to get them into logical trouble. We urge readers to carefully avoid the use of the phrases *one-to-one* and *onto* as well as avoid all the intuitive definitions you may have heard for those phrases. Instead, we recommend you to use only the terms *injective* and *surjective* and their formal definitions.

Proposition A.2.9. Given functions $f : X \rightarrow Y$ and $g : Y \rightarrow Z$, the following hold:

- (i) If f and g are both injective, then $g \circ f$ is injective.
- (ii) If f and g are both surjective, then $g \circ f$ is surjective.
- (iii) If f and g are both bijective, then $g \circ f$ is bijective.

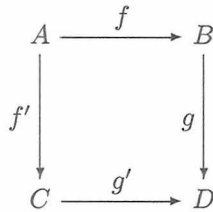
- (iv) If $g \circ f$ is injective, then f is injective.
- (v) If $g \circ f$ is surjective, then g is surjective.
- (vi) If $g \circ f$ is bijective, then g is surjective and f is injective.

Proof.

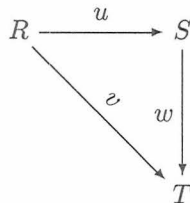
- (i) If $x, x' \in X$ and $g(f(x)) = (g \circ f)(x) = (g \circ f)(x') = g(f(x'))$, then since g is injective, we have that $f(x) = f(x')$, but since f is injective, we must have $x = x'$.
- (ii) For any $z \in Z$, because g is surjective, there is a $y \in Y$ such that $g(y) = z$. Because f is surjective, there is an $x \in X$ such that $f(x) = y$, which implies that $(g \circ f)(x) = g(f(x)) = g(y) = z$.
- (iii) Follows immediately from (i) and (ii).
- (iv) If $x, x' \in X$ and $f(x) = f(x')$, then $(g \circ f)(x) = g(f(x)) = g(f(x')) = (g \circ f)(x')$. Since $g \circ f$ is injective, we must have that $x = x'$.
- (v) If $z \in Z$, because $g \circ f$ is surjective, there exists an $x \in X$ such that $g(f(x)) = (g \circ f)(x) = z$, so letting $y = f(x) \in Y$, we have $g(y) = z$.
- (vi) Follows immediately from (iv) and (v). $\square$

To conclude this section, we describe a useful and common form of notation called a *commutative diagram*.

Definition A.2.10. We say that the diagram



commutes if $g \circ f = g' \circ f'$. Similarly, we say that the diagram



commutes if $w \circ u = v$.

A.2.3 Products and Projections

When we defined Cartesian products of sets in Definition A.1.10(viii), we only did it for finite lists of sets. We can now define infinite products.

Definition A.2.11. Given any set I (called the index set), and given any list of sets $\mathcal{A} = [A_\alpha \mid \alpha \in I]$ indexed by I , we define the Cartesian product of the sets in $\mathcal{A}$ to be the set

$$\prod_{\alpha \in I} A_\alpha = \left\{ \phi : I \rightarrow \bigcup_{\alpha \in I} A_\alpha \mid \phi(\alpha) \in A_\alpha \text{ for all } \alpha \in I \right\}.$$

Remark A.2.12. Note that the list $\mathcal{A}$ is not necessarily a set itself, because we permit duplicate elements. That is, we wish to allow constructions like $S \times S$, which corresponds to the list $[S, S]$, whereas as a set this list would have only one element $\{S, S\} = \{S\}$.

Example A.2.13.

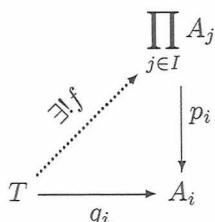
- (i) If the index set is $I = \{1, 2\}$, then $\prod_{i \in I} A_i = \{f \mid \{1, 2\} \rightarrow A_1 \cup A_2 \text{ s.t. } f(i) \in A_i\} = \{(a_1, a_2) \mid a_1 \in A_1, a_2 \in A_2\} = A_1 \times A_2$, so this definition is equivalent to our earlier definition in the case that the index set has only two elements.
- (ii) Following the outline of the previous example, one can show that if the index set is finite, then the new definition of a product is equivalent to our earlier definition for a Cartesian product of a finite number of sets.
- (iii) For a nonempty indexing set I , if there exists $\alpha \in I$ such that $A_\alpha = \emptyset$, then $\prod_{\beta \in I} A_\beta = \emptyset$.

Cartesian products have natural maps called *projections* mapping out of them.

Definition A.2.14. Given sets A and B , define the projections $p_1 : A \times B \rightarrow A$ and $p_2 : A \times B \rightarrow B$ by $p_1(a, b) = a$ and $p_2(a, b) = b$.

More generally, given any ordered list $\mathcal{A} = [A_i \mid i \in I]$ of sets indexed by a set I , the i th projection p_i is a map $\prod_{i \in I} A_i \rightarrow A_i$ defined by $p_i(\phi) = \phi(i)$.

Proposition A.2.15. The product is universal in the following sense: given any other set T with functions $q_i : T \rightarrow A_i$ for every $i \in I$, there exists a unique function $f : T \rightarrow \prod_{j \in I} A_j$ such that for every $i \in I$ we have $q_i = p_i \circ f$.



Note that the same map f makes the above diagram commute for all $i \in I$.

Proof. For each $t \in T$, let $f(t) \in \prod_{j \in I} A_j$ be given by $f(t)(j) = q_j(t) \in A_j$. Clearly $f(t) \in \prod_{j \in I} A_j$. Moreover, $(p_i \circ f)(t) = f(t)(i) = q_i(t)$, as desired. Finally, to show uniqueness, note that if there is another $g : T \rightarrow \prod_{j \in I} A_j$ with $(p_i \circ g)(t) = q_i(t)$ for every $i \in I$ and every $t \in T$, then for each $t \in T$ and for every $i \in I$, we have $g(t)(i) = q_i(t) = f(t)(i)$; so $g(t) = f(t)$ for every $t \in T$. Hence $g = f$. $\square$

A.2.4 Pitfall: Well-Defined Functions

There are some potential pitfalls to defining functions—it is very easy to write down a rule that seems like it should be a function, but which is ambiguous in some way or other. For example, if we define $q : \mathbb{Q} \times \mathbb{Q} \times \mathbb{Q} \rightarrow \mathbb{Q}$ to be the rule $f(a, b, c) = a/b/c$, it is unclear what is meant, because division is a binary operation—division takes only two inputs at a time. The expression $a/b/c$ could mean either $(a/b)/c$ or $a/(b/c)$, and since those expressions give different results, f is not a function. It is traditional to say that the function f is “not well defined.” This phrase is an abuse of language, because f is not a function at all, but everyone uses this phrase instead of the more correct “ f is not a function.”

This pitfall also occurs when we try to define functions from the set of equivalence classes on a set. It is often convenient to use specific representatives of the equivalence class to define the function. But this can lead to ambiguity which prevents the supposed function from being well defined (i.e., from being a function).

For example, consider the partition of $\mathbb{Z}$ into the equivalence classes of $\mathbb{Z}_4$, as in Example A.1.18(i). We have four equivalence classes:

$$\begin{aligned} [0]_4 &= \{\dots, -4, 0, 4, 8, \dots\}, \\ [1]_4 &= \{\dots, -7, -3, 1, 5, \dots\}, \\ [2]_4 &= \{\dots, -6, -2, 2, 6, \dots\}, \\ [3]_4 &= \{\dots, -5, -1, 3, 7, \dots\}. \end{aligned}$$

Similarly, the partition $\mathbb{Z}_3$ has three equivalence classes:

$$\begin{aligned} [0]_3 &= \{\dots, -6, -3, 0, 3, 6, \dots\}, \\ [1]_3 &= \{\dots, -5, -2, 1, 4, 7, \dots\}, \\ [2]_3 &= \{\dots, -4, -1, 2, 5, 8, \dots\}. \end{aligned}$$

We might try to define a function $g : \mathbb{Z}_4 \rightarrow \mathbb{Z}_3$ by $g([a]_4) = [a]_3$, so we should have

$$g([0]_4) = [0]_3,$$

but also $[0]_4 = [4]_4$, so

$$g([4]_4) = [4]_3 = [1]_3 \neq [0]_3 = g([0]_4).$$

So the rule for g is ambiguous—it is unclear which of the possible values for $g([0]_4)$ the rule is specifying. To be a function, it must unambiguously identify a unique choice for each input. Therefore, g is not a function. Again we abuse language and say “the function g is not well defined.”

On the other hand, we may define a function $f : \mathbb{Z}_4 \rightarrow \mathbb{Z}_2$ by $f([a]_4) = [a]_2$, which gives

$$\begin{aligned} f([0]_4) &= [0]_2 = \{\text{Even integers}\}, \\ f([1]_4) &= [1]_2 = \{\text{Odd integers}\}, \\ f([2]_4) &= [2]_2 = \{\text{Even integers}\}, \\ f([3]_4) &= [3]_2 = \{\text{Odd integers}\}. \end{aligned}$$

In this case f is well defined, because any time $[a] = [a']$ we have $a' \equiv a \pmod{4}$, which implies that $4|(a' - a)$ and hence $2|(a' - a)$, so the equivalence classes $\pmod{2}$ are the same: $[a]_2 = [a']_2$.

Whenever we wish to define a function f whose domain is a set of equivalence classes, if the definition is given in terms of specific representatives of the equivalence classes, then we must check that the function is well defined. That means we must check that using two different representatives a and a' of the same equivalence class $[a] = [a']$ gives $f([a]) = f([a'])$.

Example A.2.16.

- (i) For any $n \in \mathbb{Z}$, define a function $\boxplus : \mathbb{Z}_n \times \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ by $[a] \boxplus [b] = [a + b]$. To check that it is well defined, we must check that if $[a] = [a']$ and $[b] = [b']$, then $[a + b] = [a' + b']$. We have $a' = a + kn$ and $b' = b + \ell n$ for some $k, \ell \in \mathbb{Z}$, so $a' + b' = a + b + kn + \ell n$, and thus $[a' + b'] = [a + b]$, as required.
- (ii) For any $n \in \mathbb{Z}$, define a function $\boxtimes : \mathbb{Z}_n \times \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ by $[a] \boxtimes [b] = [ab]$. Again, to check that it is well defined, we must check that if $[a] = [a']$ and $[b] = [b']$, then $[ab] = [a'b']$. We have $a' = a + kn$ and $b' = b + \ell n$ for some $k, \ell \in \mathbb{Z}$, so $a'b' = (a + kn)(b + \ell n) = ab + n(bk + a\ell + k\ell)$, and thus $[a'b'] = [ab]$, as required.
- (iii) Given $n, m \in \mathbb{Z}$, try to define a function $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_m$ by $f([a]_n) = [a]_m$. If m does not divide n , then this function is not well defined because $[0]_n = [n]_n$, but $f([0]_n) = [0]_m \neq [n]_m = f([n]_n)$ because $m \nmid (n - 0)$.
- (iv) The previous example *does* give a well-defined function if m divides n because $f([a + kn]_n) = [a + kn]_m = [a]_m = f([a]_n)$.
- (v) If F is the set of equivalence classes of fractions, as in Example A.1.13(v), then we define a function $p : F \times F \rightarrow F$ by $p(a/b, c/d) = (ad + bc)/(bd)$. This is the function we usually call “+” for fractions. To check that it is well defined, we must check that for any $a'/b' \sim a/b$ and $c'/d' \sim c/d$ we have $p(a'/b', c'/d') \sim p(a/b, c/d)$. Although this is not immediately obvious, it requires only basic arithmetic to check, and we leave the computation to the reader.

Remark A.2.17. If a rule is already known to be a function, or if it is obviously unambiguously defined, then you need not check that it is well defined. In particular, if a function is the composition of two other (well-defined) functions, then it is already known to be a function (see Proposition A.2.6), and therefore it is already well defined.

A.2.5 Inverses

Definition A.2.18. Let $f : X \rightarrow Y$ be a function.

- (i) The function f is right invertible if there exists a right inverse, that is, if there exists $g : Y \rightarrow X$ such that $f \circ g = I_Y$, where $I_Y : Y \rightarrow Y$ denotes the identity map on Y .
- (ii) The function f is left invertible if there exists a left inverse, that is, if there exists $g : Y \rightarrow X$ such that $g \circ f = I_X$, where $I_X : X \rightarrow X$ denotes the identity map on X .

Theorem A.2.19. Let $f : X \rightarrow Y$ be given.

- (i) The function f is right invertible if and only if it is surjective.
- (ii) The function f is left invertible if and only if it is injective.
- (iii) The function f is bijective (both surjective and injective) if and only if there is a function f^{-1} which is both a left inverse and a right inverse. In this case, the function f^{-1} is the unique inverse, and we say that f is an invertible function.

Proof.

- (i) ($\implies$) Let $y \in Y$, and let $g : Y \rightarrow X$ be a right inverse of f . Define $x = g(y)$. This implies $f(x) = f(g(y)) = y$. Hence, f is surjective.
 ($\impliedby$) Since f is surjective, each set in $\mathcal{B} = \{f^{-1}(y)\}_{y \in Y}$ is nonempty. By the axiom of choice, there exists $\phi : \mathcal{B} \rightarrow \cup_{y \in Y} f^{-1}(y) = X$ such that $\phi(f^{-1}(y)) \in f^{-1}(y)$ for each $y \in Y$. Define $g : Y \rightarrow X$ as $g(y) = \phi(f^{-1}(y))$, so that $f(g(y)) = y$ for each $y \in Y$.
- (ii) ($\implies$) If $f(x_1) = f(x_2)$, then $g(f(x_1)) = g(f(x_2))$, which implies that $x_1 = x_2$. Hence, f is injective.
 ($\impliedby$) Choose $x_0 \in X$. Define $g : Y \rightarrow X$ by

$$g(y) = \begin{cases} x & \text{if there exists } x \text{ such that } f(x) = y, \\ x_0 & \text{otherwise.} \end{cases}$$

This is well defined since f is injective ($f(x_1) = f(x_2)$ implies $x_1 = x_2$). Thus, $g(f(x)) = x$ for each $x \in X$.

- (iii) ($\implies$) Since f is bijective, there exists both a left inverse and a right inverse. Let f^{-L} be any left inverse, and let f^{-R} be any right inverse. For all $y \in Y$

we have $f^{-L}(y) = (f^{-L} \circ (f \circ f^{-R}))(y) = ((f^{-L} \circ f) \circ f^{-R})(y) = f^{-R}(y)$, so every left inverse is equal to every right inverse.

($\Leftarrow$) Since f^{-1} is a left inverse, f is injective, and since f^{-1} is a right inverse, f is surjective. $\square$

Corollary A.2.20. *Let $f : X \rightarrow Y$ and $g : Y \rightarrow X$ be given.*

- (i) *If g is a bijection and $g(f(x)) = x$ for all $x \in X$, then $f = g^{-1}$.*
- (ii) *If g is a bijection and $f(g(y)) = y$ for all $y \in Y$, then $f = g^{-1}$.*

Proof.

- (i) Since g is a bijection, we have that $g^{-1} : X \rightarrow Y$ exists, and thus, $g^{-1}(g(y)) = y$ for all $y \in Y$. Thus, for all $x \in X$, we have that $g^{-1}(g(f(x))) = f(x)$ and $g^{-1}(g(f(x))) = g^{-1}(x)$.
- (ii) Since g is a bijection, we have that $g^{-1} : X \rightarrow Y$ exists, and thus, $g(g^{-1}(x)) = x$ for all $x \in X$. Thus, for all $x \in X$, we have that $f(g(g^{-1}(x))) = f(x)$ and $f(g(g^{-1}(x))) = g^{-1}(x)$. $\square$

Corollary A.2.21. *Let $f : X \rightarrow Y$. If f is invertible, then $(f^{-1})^{-1} = f$.*

Proof. Since f^{-1} exists, we have that $f^{-1}(f(x)) = x$ for all $x \in X$ and $f(f^{-1}(y)) = y$ for all $y \in Y$. Thus, f is the left and right inverse of f^{-1} , which implies that $(f^{-1})^{-1}$ exists and equals f . $\square$

Example A.2.22. The fundamental theorem of calculus says that the map $\text{Int} : C^{n-1}([a, b]; \mathbb{F}) \rightarrow C^n([a, b]; \mathbb{F})$ given by $\text{Int } f = \int_a^x f(t) dt$ is a right inverse of the map $D = \frac{d}{dx} : C^n([a, b]; \mathbb{F}) \rightarrow C^{n-1}([a, b]; \mathbb{F})$ because $D \circ \text{Int} = \text{Id}_{C^{n-1}}$ is the identity. This shows that D is surjective and Int is injective. Because D is not injective, it does not have a left inverse. But it has infinitely many right inverses: for any constant $C \in \mathbb{F}$ the map $\text{Int} + C$ is also a right inverse to D .

A.3 Orderings

A.3.1 Total Orderings

Definition A.3.1. *A relation ($<$) on the set S is a total ordering (and the set is called totally ordered) if for all $x, y, z \in S$, we have the following:*

- (i) *Trichotomy: If $x, y \in S$, then exactly one of the following relations holds:*

$$x < y, \quad x = y, \quad y < x.$$

- (ii) *Transitivity: If $x < y$ and $y < z$, then $x < z$.*

The relation $x \leq y$ means that $x < y$ or $x = y$ and is the negation of $y < x$. A total ordering is also often called a linear ordering.

Definition A.3.2. A set S with a total ordering $<$ is well ordered if every nonempty subset $X \subset S$ has a least element; that is, an element $x \in X$ such that $x \leq y$ for every $y \in X$.

We take the following as an axiom of the natural numbers.

Axiom A.3.3 (Well-Ordering Axiom for Natural Numbers). The natural numbers with the usual ordering $<$ are well ordered.

The well-ordering axiom (WOA) allows us to speak of the first element of a set that does, or does not, satisfy some property.

Example A.3.4. A prime number is an integer greater than 1 that has no divisors other than itself and 1. We can use the WOA to show that every positive integer greater than 1 has a prime divisor. Let

$$S = \{n \in \mathbb{Z} \mid n > 1 \text{ and } n \text{ has no prime divisor}\}.$$

By the WOA, if S is not empty, there is a least element $s \in S$. If s is prime, then s divides itself, and the claim is satisfied. If s is not prime, then it must have a divisor d with $1 < d < s$. Since $d \notin S$, there is a prime p that divides d , but now p also divides s , since p divides a divisor of s . This is a contradiction, so S must be empty.

Remark A.3.5. Note that the set $\mathbb{Z}$ of all integers is not well ordered, since it does not have a least element. However, for any $n \in \mathbb{Z}$, the set $\mathbb{Z}^{>n} = \{x \in \mathbb{Z} \mid x > n\}$ is well ordered, since we can define a bijective map $f : \mathbb{Z}^{>n} \rightarrow \mathbb{N}$ to the natural numbers $f(x) = x - n - 1$, which preserves ordering. In particular, the positive integers $\mathbb{Z}^{>0} = \mathbb{Z}^+$ are well ordered.

A.3.2 Induction

The principle of mathematical induction provides an extremely powerful way to prove many theorems. It follows from the WOA of the natural numbers.

Theorem A.3.6 (Principle of Mathematical Induction). Given any property P , in order to prove that P holds for all positive integers numbers, it suffices to prove that

- (i) P holds for 1;
- (ii) for every $k \in \mathbb{Z}^+$, if P holds for k , then P holds for $k + 1$.

The first step— P holds for 1—is called the *base case*, and the second step is called the *inductive step*. The statement P holds for the number k is called the *induction hypothesis*.

Proof. Assume that both the base case and the inductive step have been verified for the property P . Let $S = \{x \in \mathbb{Z}^+ \mid P \text{ does not hold for } x\}$ be the set of all positive integers for which the property P does not hold. If S is not empty, then by the WOA it must have a least element $s \in S$. By the base case of the induction, we know that $s > 1$. Let $k = s - 1$. Since $k < s$, we have that P holds for k . Since $k > 0$, the inductive step implies that P also holds for $k + 1 = s$, a contradiction. Therefore S is empty, and P holds for all $n \in \mathbb{Z}^+$. $\square$

Example A.3.7.

- (i) We use induction to prove for any positive integer n that $1 + 3 + \cdots + (2n - 1) = n^2$. The base case $(2 \cdot 1 - 1) = 1 = 1^2$ is immediate. Now, given the induction hypothesis that $1 + 3 + \cdots + (2k - 1) = k^2$, we have $1 + 3 + \cdots + (2k - 1) + (2(k + 1) - 1) = k^2 + (2(k + 1) - 1) = k^2 + 2k + 1 = (k + 1)^2$, as required. So the claim holds for all positive integers n .
- (ii) We use induction to prove for any integer $n > 0$ that $\sum_{i=1}^n i = \frac{(n+1)n}{2}$. The base case $\sum_{i=1}^1 i = 1$ is immediate. Now, given the inductive hypothesis that $\sum_{i=1}^k i = \frac{(k+1)k}{2}$, we have $\sum_{i=1}^{k+1} i = (k + 1) + \sum_{i=1}^k i = (k + 1) + \frac{(k+1)k}{2} = \frac{(k+1)(k+2)}{2}$, as required. So the claim holds for all positive integers n .

Corollary A.3.8. *The principle of induction can be applied to any set of the form*

$$\mathbb{N} + a = \{x \in \mathbb{Z} \mid x \geq a\}$$

by starting the base case at a . That is, a property P holds for all of $\mathbb{N} + a$ if

- (i) *P holds for a ;*
- (ii) *for every $k \in \mathbb{N} + a$, if P holds for k , then P holds for $k + 1$.*

Proof. Let $P'(n)$ be the property $P(n - 1 + a)$. Application of the usual induction principle to P' is equivalent to our new induction principle for P on T . $\square$

Definition A.3.9. *An element b in a totally ordered set S is an upper bound of the set $E \subset S$ if $x \leq b$ for all $x \in E$. Similarly, an element $a \in S$ is a lower bound of the set $E \subset S$ if $a \leq x$ for all $x \in E$. If E has an upper (respectively, lower) bound, we say that E is bounded above (respectively, below).*

If b is an upper bound for $E \subset S$ that has the additional property that $c < b$ implies that c is not an upper bound of E , then b is called the least upper bound of E or the supremum of E and is denoted $\sup E$.

Similarly, if a is a lower bound of E such that $a < d$ implies that d is not a lower bound of E , then we call a the greatest lower bound or the infimum of E and denote it by $\inf E$.

Both $\inf E$ and $\sup E$ are unique, if they exist.

Definition A.3.10. An ordered set is said to have the least upper bound property, or l.u.b. property, if every nonempty subset that is bounded above has a least upper bound (supremum).

Example A.3.11.

- (i) The set $\mathbb{Z}$ of integers has the l.u.b. property. To see this, let $S \subset \mathbb{Z}$ be any subset that is bounded above, and let $T = \{x \in \mathbb{Z} \mid x > s \forall s \in S\}$ be the set of all upper bounds for S . Because T is not empty (S is bounded), the WOA guarantees that T has a least element t . The element t is an l.u.b. for S .
- (ii) The set $\mathbb{Q}$ of rational numbers does *not* have the l.u.b. property; for example, the set $E = \{x \in \mathbb{Q} \mid x^2 < 2\} \subset \mathbb{Q}$ does not have an l.u.b. in $\mathbb{Q}$ —but the l.u.b. for E does exist in $\mathbb{R}$, namely, $\sqrt{2}$.
- (iii) The real numbers $\mathbb{R}$ have the l.u.b. property. We take this as an axiom of the real numbers.

Theorem A.3.12. Let S be an ordered set with the l.u.b. property. If a nonempty set $E \subset S$ is bounded below, then its infimum exists in S .

Proof. Assume that $E \subset S$ is nonempty and bounded below. Let L denote the set of all lower bounds of E , which is nonempty by hypothesis. Since E is nonempty, then any given $x \in E$ is an upper bound for L , and thus $a = \sup L$ exists in S . We claim that $a = \inf E$. Suppose not. Then either a is not a lower bound of E or a is a lower bound, but not the greatest lower bound. If the former is true, then there exists $x \in E$ such that $x < a$, but this contradicts a as the l.u.b. of L since x would also be an upper bound of L . If the latter is true, then there exists $c \in S$ such that $a < c$ and c is a lower bound of E . This is a contradiction because $c \in L$, and thus a cannot be the l.u.b. of L . Thus, $a = \inf E$. $\square$

A.3.3 Partial Orderings

Definition A.3.13. A partial ordering of a nonempty set A is a relation ($\leq$) on A that is reflexive ($x \leq x$ for all $x \in A$), transitive ($x \leq y$ and $y \leq z \Rightarrow x \leq z$), and antisymmetric ($x \leq y$ and $y \leq x \Rightarrow x = y$). A partially ordered set (or poset, for short) is a set with a partial order.

Remark A.3.14. Two elements x, y of a partially ordered set are comparable if $x \leq y$ or $y \leq x$. A partial ordering does not require every pair of elements to be comparable, but in the special case that every pair of elements is comparable, then $\leq$ defines a total (linear) ordering; see Section A.3.1.

Definition A.3.15. Let $(X, \leq)$ be a partially ordered set. If for some $a \in X$ we have $x \leq a$ for every x that is comparable to a , we say a is a maximal element of X . If $x \leq a$ for all $x \in X$, we say that a is the greatest element.

Similarly, if for some $b \in X$ we have $b \leq x$ for every x that is comparable to b , we say b is a minimal element of X . If $b \leq x$ for all $x \in X$, we say that b is the least element.

A.4 Zorn's Lemma, the Axiom of Choice, and Well Ordering

We saw in Example A.2.13(iii) that if $A_\alpha = \emptyset$ for some $\alpha \in I$, then $\prod_{\beta \in I} A_\beta = \emptyset$. The converse states that if $A_\alpha \neq \emptyset$ for all $\alpha \in I$, then $\prod_{\alpha \in I} A_\alpha \neq \emptyset$. This statement, however, is independent of the regular axioms of set theory, yet it stands as one of the most fundamental ideas of mathematics. It is called the axiom of choice.

Theorem A.4.1 (Axiom of Choice). *Let I be a set, and let $\mathcal{A} = \{X_\alpha\}_{\alpha \in I}$ be a set of nonempty sets indexed by I . There is a function $f : I \rightarrow \bigcup_{\alpha \in I} X_\alpha$ such that $f(\alpha) \in X_\alpha$ for all $\alpha \in I$.*

The axiom of choice has two other useful equivalent formulations—the *well-ordering principle* and *Zorn's lemma*.

Theorem A.4.2 (Well-Ordering Principle). *If A is a nonempty set, then there exists a total (linear) ordering $\leq$ of A such that $(A, \leq)$ is well ordered.*

Remark A.4.3. It is important to note that the choice of ordering may well differ from the natural order on the set. For example, $\mathbb{R}$ has a natural choice of total order on it, but it is *not* true that $\mathbb{R}$ with the usual ordering is well ordered—the subset $(0, 1)$ has no least element. The ordering that makes $\mathbb{R}$ well ordered is entirely different, and it is not at all obvious what that ordering is.

The third equivalent formulation is Zorn's lemma.

Theorem A.4.4 (Zorn's Lemma). *Let $(X, \leq)$ be a partially ordered set. If every chain in X has an upper bound in X , then X contains a maximal element.*

By *chain* we mean any nonempty subset $C \subset X$ such that C is totally ordered, that is, for every $\alpha, \beta \in C$ we have either $\alpha \leq \beta$ or $\beta \leq \alpha$. A chain C is said to have an upper bound in X if there is an element $x \in X$ such that $\alpha \leq x$ for every $\alpha \in C$.

Theorem A.4.5. *The following three statements are equivalent:*

- (i) *The axiom of choice.*
- (ii) *The well-ordering principle.*
- (iii) *Zorn's lemma.*

We do not prove this theorem here, as it would take us too far afield. The interested reader may consult [Hal74].

Definition A.4.6. Let $(X, \leq)$ be a well-ordered set. For each $y \in X$, define $s(y) = \{x \in X \mid x < y\}$, and define $\bar{s}(y)$ to be $s(y) \cup y$. We say that a subset $S \subseteq X$ is a segment of X if either $S = X$ or there exists $y \in X$ such that $S = s(y)$. If the complement $s(y)^c = X \setminus s(y)$ is nonempty, we call the least element of $s(y)^c$ the immediate successor of y .

Example A.4.7.

- (i) If z is the immediate successor of y , then we have $s(z) = \bar{s}(y)$.
- (ii) Every element of X except the greatest element (if it exists) has an immediate successor.

A.5 Cardinality

It is often useful to think about the size of a set—we call this the *cardinality* of the set. In the case that the set is finite, the cardinality is simply the number of elements in the set. In the case that the set is infinite, cardinality is more subtle.

Definition A.5.1. We say that two sets A and B have the same cardinality if there exists a bijection $f : A \rightarrow B$.

Definition A.5.2. It is traditional to use the symbol $\aleph_0$ to denote the cardinality of the integers. A set with this cardinality is countable. Any set with cardinality strictly greater than $\aleph_0$ is called uncountable. If a set A has the same cardinality as the set $\{1, 2, \dots, n\}$, we say that the cardinality of A is n . We often write $|A|$ to denote the cardinality of A , so

$$|\mathbb{Z}| = \aleph_0 \quad \text{and} \quad |\{1, 2, \dots, n\}| = n.$$

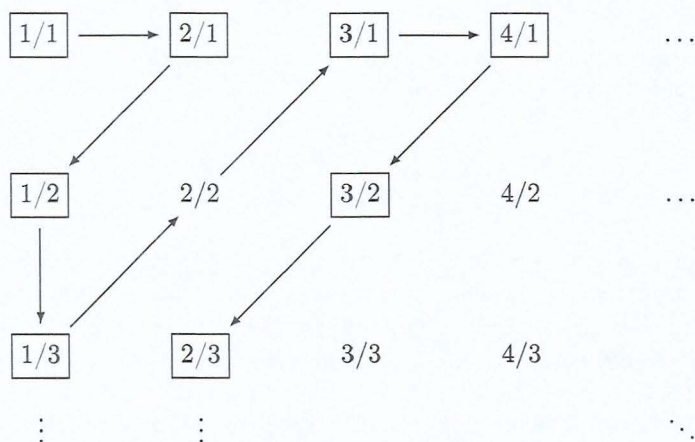
Example A.5.3.

- (i) The cardinality of the even integers $2\mathbb{Z}$ is the same as that of $\mathbb{Z}$. This follows from the bijection $g : \mathbb{Z} \rightarrow 2\mathbb{Z}$ given by $g(n) = 2n$, which is bijective because it has a (left and right) inverse $g^{-1} : 2\mathbb{Z} \rightarrow \mathbb{Z}$ given by $g^{-1}(m) = \frac{m}{2}$.
- (ii) The cardinality of the set $\mathbb{Z}^+$ of positive integers is the same as the cardinality of the set of integers (it has cardinality $\aleph_0$). To see this, use the function $h : \mathbb{Z}^+ \rightarrow \mathbb{Z}$ given by $h(2n) = n$ and $h(2n+1) = -n$. This function is clearly bijective, so the cardinality of $\mathbb{Z}^+$ is $\aleph_0$.

- (iii) The set $\mathbb{Q}^+$ of positive rational numbers is countable. To prove this, we show that it has the same cardinality as the positive integers $\mathbb{Z}^+$. This can be seen from the following argument. List all the positive rationals in the lower right quadrant of the plane, as follows:

$1/1$	$2/1$	$3/1$	$4/1$	...
$1/2$	$2/2$	$3/2$	$4/2$	...
$1/3$	$2/3$	$3/3$	$4/3$	...
$\vdots$	$\vdots$			$\ddots$

Now number the rationals by proceeding along the following zig-zag route, skipping any that have already been numbered (e.g., $2/2$ was counted as $1/1$, so it is not counted again):



In this way we get a function $\phi : \mathbb{Q}^+ \rightarrow \mathbb{Z}^+$, with $\phi(1/1) = 1$, $\phi(2/1) = 2$, $\phi(1/2) = 3$, $\phi(1/3) = 4$, and so forth. This function is clearly both injective and surjective, so the cardinality of $\mathbb{Q}^+$ is the same as $\mathbb{Z}^+$, and hence $\mathbb{Q}^+$ is countable.

- (iv) We do not prove it here, but it is not difficult to show that any subset of $\mathbb{Q}$ is either finite or countably infinite.

- (v) The real numbers in the interval $(0, 1)$ are not countable, as can be seen from the following sketch of *Cantor's diagonal argument*. If the set $(0, 1)$ were countable, there would have to exist a bijection $c : \mathbb{Z}^+ \rightarrow (0, 1)$. Listing the decimal expansion of each real number, in order, we have

$$\begin{aligned} c(1) &= 0.a_{11}a_{12}a_{13}\dots, \\ c(2) &= 0.a_{21}a_{22}a_{23}\dots, \\ c(3) &= 0.a_{31}a_{32}a_{33}\dots, \\ &\vdots \quad \vdots \end{aligned}$$

where a_{ij} is the j th digit of the decimal expansion of $c(i)$.

Now consider the real number b with decimal expansion $0.b_1b_2b_3\dots$, where for each positive integer k the digit b_k is chosen to be some digit in the range $\{1, 2, \dots, 8\}$ which is different from a_{kk} . We have $b \in (0, 1)$, but for any $\ell \in \mathbb{Z}^+$ we have that $b \neq c(\ell)$ because their decimal expansions differ in the ℓ th digit:^a $b_\ell \neq a_{\ell\ell}$. Thus the map c is not surjective, a contradiction.

^aThis last point is not completely obvious—some care is needed to prove that the different decimal expansions give different real numbers. For more details on this argument see, for example, [Hal74].

Proposition A.5.4. *If $|B| = n \in \mathbb{Z}^+$ and $b \in B$, then $|B \setminus b| = n - 1$.*

Proof. The expression $|B| = n$ means that there is a bijection $\phi : B \rightarrow \{1, 2, \dots, n\}$. We define a new bijection $\psi : (B \setminus b) \rightarrow \{1, 2, \dots, n - 1\}$ by

$$\psi(x) = \begin{cases} \phi(x) & \text{if } \phi(x) \neq n, \\ \phi(b) & \text{if } \phi(x) = n. \end{cases}$$

We claim that ψ is a bijective map onto $\{1, 2, \dots, n - 1\}$. First, for any $x \in (B \setminus b)$, if $\phi(x) \neq n$, then $\psi(x) \neq n$, while if $\phi(x) = n$, then since $x \neq b$, we have $\psi(x) = \phi(b) \neq \phi(x) = n$. So ψ is a well-defined map $\psi : (B \setminus b) \rightarrow \{1, 2, \dots, n - 1\}$.

For injectivity, consider $x, y \in (B \setminus b)$ with $\psi(x) = \psi(y)$. If $\phi(x) \neq n$ and $\phi(y) \neq n$, then $\phi(x) = \psi(x) = \psi(y) = \phi(y)$, so by injectivity of ϕ , we have $x = y$. If $\phi(x) \neq n$ but $\phi(y) = n$, then $\phi(x) = \psi(x) = \psi(y) = \phi(b)$, so $x = b$, a contradiction. Finally, if $\phi(x) = n = \phi(y)$, then by injectivity of ϕ we have $x = y$.

For surjectivity, consider any $k \in \{1, 2, \dots, n - 1\}$. By surjectivity of ϕ there exists an $x \in B$ such that $\phi(x) = k$. If $x \neq b$, then since $k \neq n$, we have $\psi(x) = \phi(x) = k$. But if $x = b$, then there exists also a $y \in B \setminus b$ such that $\phi(y) = n$, so $\psi(y) = \phi(b) = k$. $\square$

Proposition A.5.5. *If B is a finite set and $A \subset B$, then $|A| \leq |B|$.*

Proof. We prove this by induction on the cardinality of B . If $|B| = 0$, then $B = \emptyset$ and also $A = \emptyset$, so $|A| = |B|$. Assume now that the proposition holds for every B with $|B| < n$. If $A = B$, then $|A| = |B|$ and we are done. But if $A \neq B$, and $|B| = n$, then there exists some $b \in B \setminus A$, and therefore $A \subset B \setminus b$. By Proposition A.5.4, the set $B \setminus b$ has cardinality $n - 1$, so by the induction hypothesis, we have $|A| \leq n - 1 < |B| = n$. $\square$

Since any injective map gives a bijection onto its image, we have, as an immediate consequence, the following corollary.

Corollary A.5.6. *If A and B are two finite sets and $f : A \rightarrow B$ is injective, then $|A| \leq |B|$.*

The previous corollary inspires the following definition for comparing cardinality of arbitrary sets.

Definition A.5.7. *For any two sets A and B , we say $|A| \leq |B|$ if there exists an injection $f : A \rightarrow B$, and we say $|A| < |B|$ if there is an injection $f : A \rightarrow B$, but there is no surjection from A to B .*

Example A.5.8.

- (i) $\aleph_0 < |\mathbb{R}|$ because there is an obvious injection $\mathbb{Z} \rightarrow \mathbb{R}$, but by Example A.5.3(v), there is no surjection $\mathbb{Z}^+ \rightarrow (0, 1)$ and hence no surjection $\mathbb{Z} \rightarrow \mathbb{R}$.
- (ii) For every nonnegative integer n we have $n < \aleph_0$ because there is an obvious injection $\{1, 2, \dots, n\} \rightarrow \mathbb{Z}$, but there is no surjection $\{1, 2, \dots, n\} \rightarrow \mathbb{Z}$.

Proposition A.5.9. *If $g : A \rightarrow B$ is a surjection, then $|A| \geq |B|$.*

Proof. By surjectivity, for each $b \in B$, there exists an $a \in A$ such that $g(a) = b$. Choose one such a for each b and call it $f(b)$. The assignment $b \mapsto f(b)$ defines a function $f : B \rightarrow A$. Moreover, if $b, b' \in B$ are such that $f(b) = f(b')$, then $b = g(f(b)) = g(f(b')) = b'$, so f is injective. This implies that $|B| \leq |A|$. $\square$

We conclude with two corollaries that follow immediately from our previous discussion, but which are extremely useful.

Corollary A.5.10. *If A and B are finite sets of the same cardinality, then any map $f : A \rightarrow B$ is injective if and only if it is also surjective.*

Remark A.5.11. The previous corollary is *not* true for infinite sets, as can be seen from the case of $2\mathbb{Z} \subset \mathbb{Z}$ and also $\mathbb{Z} \subset \mathbb{Q}$. The inclusions are injective maps that are not surjective, despite the fact that these sets all have the same cardinality.

Corollary A.5.12 (Pigeonhole Principle). *If $|A| > |B|$, then given any function $f : A \rightarrow B$, there must be at least two elements $a, a' \in A$ such that $f(a) = f(a')$.*

This last corollary gets its name from the example where A is a set of pigeons and B is a set of pigeonholes. If there are more pigeons than pigeonholes, then at least two pigeons must share a pigeonhole. This “principle” is used in many counting arguments.

B

The Complex Numbers and Other Fields

For every complex problem there is an answer that is clear, simple, and wrong.
—H. L. Mencken

In this appendix we briefly review the fundamental properties of the field of complex numbers and also general fields.

B.1 Complex Numbers

B.1.1 Basics of Complex Numbers

Definition B.1.1. Let i be a formal symbol (representing a square root of -1). Let $\mathbb{C}$ denote the set

$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}.$$

Elements of $\mathbb{C}$ are called complex numbers. We define addition of complex numbers by

$$(a + bi) + (c + di) = (a + c) + (b + d)i,$$

and we define multiplication of complex numbers by

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$

Example B.1.2. Complex numbers of the form $a + 0i$ are usually written just as a , and those of the form $0 + bi$ are usually written just as bi .

We verify that i has the expected property

$$i^2 = (0 + 1i)^2 = (0 + 1i)(0 + 1i) = (0 - 1) + (0 + 0)i = -1.$$

Definition B.1.3. For any $z = a + bi \in \mathbb{C}$ we define the complex conjugate of z to be $\bar{z} = a - bi$, and we define the modulus (sometimes also called the norm) of z to be $|z| = \sqrt{z\bar{z}} = \sqrt{a^2 + b^2}$. We also define the real part $\Re(z) = a$ and the imaginary part $\Im(z) = b$.

Note that $|z| \in \mathbb{R}$ for any $z \in \mathbb{C}$.

Proposition B.1.4. Addition and multiplication of complex numbers satisfy the following properties for any $z, w, v \in \mathbb{C}$:

- (i) Associativity of addition: $(v + w) + z = v + (w + z)$.
- (ii) Commutativity of addition: $z + w = w + z$.
- (iii) Associativity of multiplication: $(vw)z = v(wz)$.
- (iv) Commutativity of multiplication: $zw = wz$.
- (v) Distributivity: $v(w + z) = vw + vz$.
- (vi) Additive identity: $0 + z = z = z + 0$.
- (vii) Multiplicative identity: $1 \cdot z = z = z \cdot 1$.
- (viii) Additive inverses: if $z = a + bi$, then $-z = -a - bi$ satisfies $z + (-z) = 0$.
- (ix) Multiplicative inverses: If $z = a + bi \neq 0$, then $|z|^{-2} \in \mathbb{R}$ and so

$$z^{-1} = \bar{z}|z|^{-2} = \frac{a - bi}{a^2 + b^2}.$$

Proof. All of the properties are straightforward algebraic manipulations. We give one example and leave the rest to the reader.

For (ix) first note that since $z \neq 0$ we have $|z|^2 = a^2 + b^2 \neq 0$, so its multiplicative inverse $(a^2 + b^2)^{-1}$ is also in $\mathbb{R}$. We have

$$z(\bar{z}|z|^{-2}) = z\bar{z}(z\bar{z})^{-1} = 1,$$

so $(\bar{z}|z|^{-2})$ is the multiplicative inverse to z . $\square$

B.1.2 Euler's Formula and Graphical Representation

Euler's Formula

For any $z \in \mathbb{C}$ we define the exponential e^z using the Taylor series

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

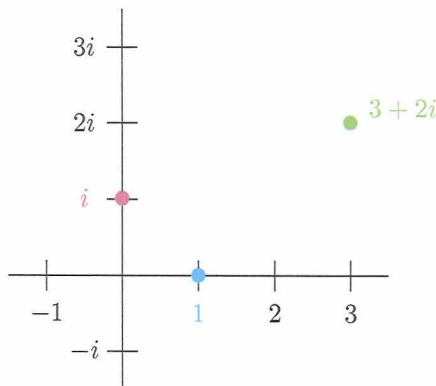


Figure B.1. A complex number $x + iy$ with $x, y \in \mathbb{R}$ is usually represented graphically in the plane as the point (x, y) . This figure shows the graphical representation of the complex numbers i , 1 , and $3 + 2i$.

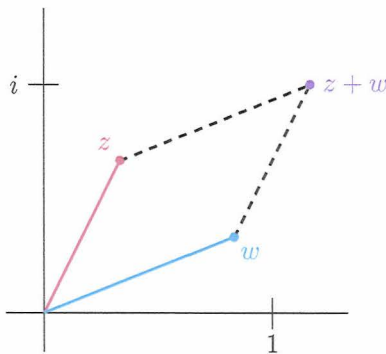


Figure B.2. Graphical representation of complex addition. Thinking of complex numbers $z = a + bi$ and $w = c + di$ as the points in the plane (a, b) and (c, d) , respectively, their sum $z + w = (a + c) + (b + d)i$ corresponds to the usual vector sum $(a, b) + (c, d) = (a + c, b + d)$ in the plane.

One of the most important identities for complex numbers is *Euler's formula* (see Proposition 11.2.12):

$$e^{it} = \cos(t) + i \sin(t) \tag{B.1}$$

for all $t \in \mathbb{C}$.

As a consequence of Euler's formula, we have *De Moivre's formula*:

$$(\cos(t) + i \sin(t))^n = (e^{it})^n = e^{int} = \cos(nt) + i \sin(nt). \tag{B.2}$$

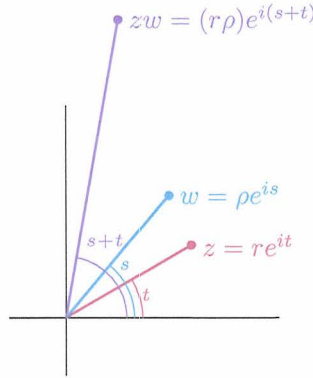


Figure B.3. Complex multiplication adds the polar angles $(s + t)$ and multiplies the moduli $(r\rho)$.

Graphical Representation

The complex numbers have a very useful graphical representation as points in the plane, where we associate the complex number $z = a + bi$ with the point $(a, b) \in \mathbb{R}^2$. See Figure B.1 for an illustration. In this representation real numbers lie along the x -axis and imaginary numbers lie along the y -axis. The modulus $|z|$ of z is the distance from the origin to z in the plane, and the complex conjugate $\bar{z}$ is the image of z under a reflection through the x -axis.

Addition of complex numbers is just the same as vector addition in the plane; so, geometrically, the complex number $z + w$ is the point in the plane corresponding to the far corner of the parallelogram whose other corners are 0, z , and w . See Figure B.2.

We can represent any point in the plane in polar form as $z = r(\cos(\theta) + i \sin(\theta))$ for some $\theta \in [0, 2\pi)$ and some $r \in \mathbb{R}$ with $r \geq 0$. Combining this with Euler's formula means that we can write every complex number in the form $z = re^{i\theta}$. In this form we have

$$|z| = |re^{i\theta}| = |r(\cos(\theta) + i \sin(\theta))| = r \quad \text{and} \quad \bar{z} = r(\cos(\theta) - i \sin(\theta)) = re^{-i\theta}.$$

We define the *sign* of $z = re^{i\theta} \in \mathbb{C}$ to be

$$\text{sign}(z) = \begin{cases} e^{i\theta} = \frac{z}{|z|} & \text{if } z \neq 0, \\ 1 & \text{if } z = 0. \end{cases} \quad (\text{B.3})$$

We can use the polar form to get a geometric interpretation of multiplication of complex numbers. If $z = re^{it}$ and $w = \rho e^{is}$, then

$$wz = r\rho e^{i(t+s)} = |z||w|(\cos(t+s) + i \sin(t+s)).$$

Multiplication of two complex numbers in polar form multiplies the moduli and adds the angles; see Figure B.3.

Similarly, $z^{-1} = \bar{z}|z|^{-2} = re^{-it}r^{-2} = r^{-1}e^{-it}$, so the multiplicative inverse changes the sign of the angle ($t \mapsto -t$) and inverts the modulus ($r \mapsto r^{-1}$). But the complex conjugate leaves the modulus unchanged and changes the sign of the angle; see Figure B.4.

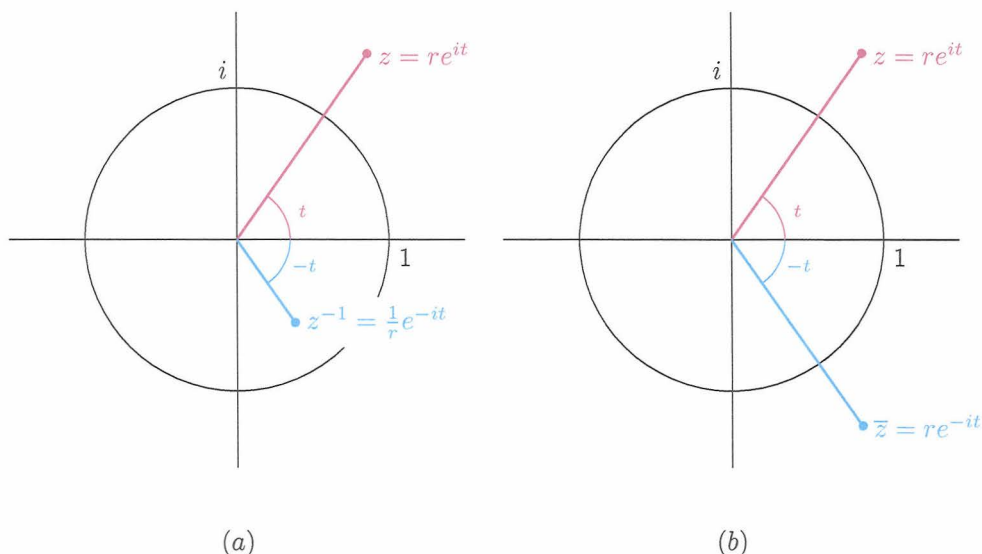


Figure B.4. Graphical representation of multiplicative inverse (a) and complex conjugate (b). The multiplicative inverse of a complex number changes the sign of the polar angle and inverts the modulus. The complex conjugate also changes the sign of the polar angle, but leaves the modulus unchanged.

B.1.3 Roots of Unity

Definition B.1.5. For $n \in \mathbb{Z}^+$ an n th root of unity is any solution to the equation $z^n = 1$ in $\mathbb{C}$. The complex number $\omega_n = e^{2\pi i/n}$ is called the primitive n th root of unity.

By the fundamental theorem of algebra (or rather its corollary, Theorem 15.3.15) there are exactly n of these n th roots of unity in $\mathbb{C}$. Euler's formula tells us that $\omega_n = \cos(2\pi/n) + i \sin(2\pi/n)$ is the point on the unit circle in the complex plane corresponding to an angle of $2\pi/n$ radians, and

$$\omega_n^k = e^{2\pi i k/n} = \cos(2\pi k/n) + i \sin(2\pi k/n).$$

See Figure B.5.

Thus we have

$$\omega_n^n = e^{2\pi i} = 1,$$

so ω_n^k is a root of unity for every $k \in \mathbb{Z}$.

If $k' \equiv k \pmod{n}$, then $k' = k + mn$ for some $m \in \mathbb{Z}$, and thus

$$\omega_n^{k'} = \omega_n^{(k+mn)} = \omega_n^k (\omega_n^n)^m = \omega_n^k.$$

The n th roots of unity are uniformly distributed around the unit circle, so their average is 0. The next proposition makes that precise.

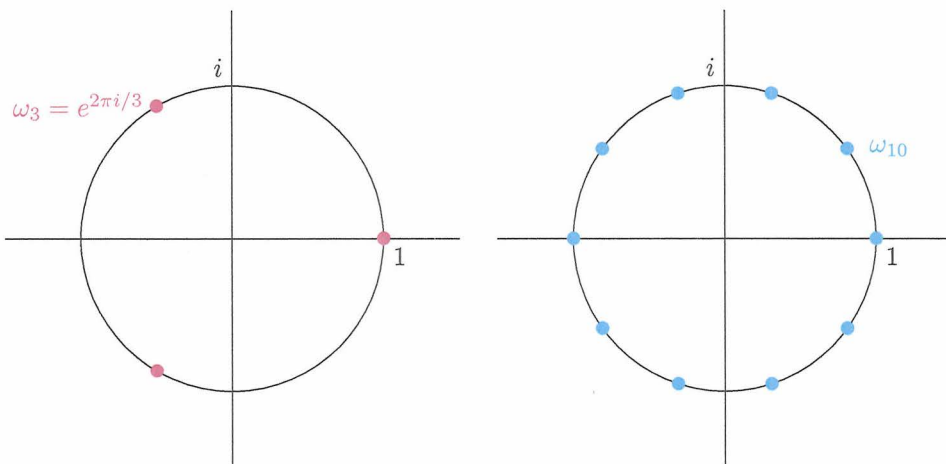


Figure B.5. Plots of all the 3rd (on the left) and 10th (on the right) roots of unity. The roots are uniformly distributed around the unit circle, so their sum is 0.

Proposition B.1.6. For any $n \in \mathbb{Z}^+$ and any $k \in \mathbb{Z}$ we have

$$\frac{1}{n} \sum_{\ell=0}^{n-1} \omega_n^{k\ell} = \frac{1}{n} \sum_{\ell=0}^{n-1} e^{2\pi i k\ell/n} = \begin{cases} 0, & k \not\equiv 0 \pmod{n}, \\ 1, & k \equiv 0 \pmod{n}. \end{cases} \quad (\text{B.4})$$

Proof. The sum $\sum_{\ell=0}^{n-1} (\omega_n^k)^\ell$ is a geometric series, so if $k \not\equiv 0 \pmod{n}$, we have

$$\sum_{\ell=0}^{n-1} \omega_n^{k\ell} = \frac{(\omega_n^k)^n - 1}{\omega_n^k - 1} = \frac{(\omega_n^n)^k - 1}{\omega_n^k - 1} = 0.$$

But if $k \equiv 0 \pmod{n}$, then

$$\frac{1}{n} \sum_{\ell=0}^{n-1} \omega_n^{k\ell} = \frac{1}{n} \sum_{\ell=0}^{n-1} 1 = 1. \quad \square$$

We conclude this section with a simple observation that turns out to be very powerful. The proof is immediate.

Proposition B.1.7. For any divisor d of n and any $k \in \mathbb{Z}$, we have

$$\omega_n^{kd} = \omega_{n/d}^k. \quad (\text{B.5})$$

Vista B.1.8. The relation (B.5) is the foundation of the fast Fourier transform (FFT). We discuss the FFT in Volume 2.

B.2 Fields

B.2.1 Axioms and Basic Properties of Fields

The real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are two examples of an important structure called a *field*. Although we only defined vector spaces with the scalars in the fields $\mathbb{R}$ or $\mathbb{C}$, you can make the same definitions for scalars from any field, and all of the results about vector spaces, linear transformations, and matrices in Chapters 1 and 2 still hold (but not necessarily the results of Chapters 3 and 4).

Definition B.2.1. A field F is a set with two binary operations, addition $(a, b) \rightarrow a + b$ and multiplication $(a, b) \rightarrow ab$, satisfying the following properties for all $a, b, c \in F$:

- (i) *Commutativity of addition:* $a + b = b + a$.
- (ii) *Associativity of addition:* $(a + b) + c = a + (b + c)$.
- (iii) *Additive identity:* There exists $0 \in F$ such that $a + 0 = a$.
- (iv) *Additive inverse:* For each $a \in F$ there exists an element denoted $(-a) \in F$ such that $a + (-a) = 0$.
- (v) *Commutativity of multiplication:* $ab = ba$.
- (vi) *Associativity of multiplication:* $(ab)c = a(bc)$.
- (vii) *Multiplicative identity:* There exists $1 \in F$ such that $1a = a$.
- (viii) *Multiplicative inverse:* For each $a \neq 0$ there exists $a^{-1} \in F$ such that $aa^{-1} = 1$.
- (ix) *Distributivity:* $a(b + c) = ab + ac$.

Example B.2.2. The most important examples of fields for our purposes are the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$. The fact that $\mathbb{C}$ is a field is the substance of Proposition B.1.4. It is easy to verify that the rational numbers $\mathbb{Q}$ also form a field.

Unexample B.2.3. The integers $\mathbb{Z}$ do not form a field, because most nonzero elements of $\mathbb{Z}$ have no multiplicative inverse. For example, the multiplicative inverse (2^{-1}) of 2 is not in $\mathbb{Z}$.

The nonnegative real numbers $[0, \infty]$ do not form a field because not all elements of $[0, \infty]$ have an additive inverse in $[0, \infty]$. For example, the additive inverse -1 of 1 is not in $[0, \infty]$.

Proposition B.2.4. *The additive and multiplicative identities of a field F are unique, as are the additive and multiplicative inverses for a given element $a \in F$. Moreover, for any $x, y, z \in F$ the following properties hold:*

- (i) $0 \cdot x = 0 = x \cdot 0$.
- (ii) If $x + y = x + z$, then $y = z$.
- (iii) If $xy = xz$ then $y = z$.
- (iv) $-(-x) = x$.
- (v) $(-y)x = -(yx) = y(-x)$.

Proof. If $a, b \in F$ and both of them are additive identities, then $a = a + b$ because b is an additive identity, but also $a + b = b$ because a is an additive identity. Therefore, we have $a = a + b = b$. The same proof applies *mutatis mutandis* to the uniqueness of the multiplicative identity.

If $b, c \in F$ are multiplicative inverses of $a \in F$, then $ca = ba = 1$, which gives $c = c(ab) = (ca)b = b$. The same proof applies *mutatis mutandis* to the uniqueness of additive inverses.

For (ii) we have that $x + y = x + z$ implies $(-x) + (x + y) = (-x) + (x + z)$. By associativity of addition, we can regroup to get $(-x + x) + y = (-x + x) + z$, and hence $y = z$. The same proof applies *mutatis mutandis* for (iii).

For (i) we have $0 \cdot x = (0 + 0) \cdot x = 0 \cdot x + 0 \cdot x$, so by (ii) we have $0 = 0 \cdot x$. Commutativity of multiplication gives $x \cdot 0 = 0$.

For (iv) we have $(-x) + (-(-x)) = 0$ by definition, but adding x on the left to both sides gives $(x + (-x)) + (-(-x)) = x$, and thus $-(-x) = x$.

Finally, for (v) we have $(-y)x + yx = (-y + y)x = 0$, and adding $-(yx)$ to both sides gives the desired result. $\square$

B.2.2 The Field $\mathbb{Z}_p$

Recall from Example A.1.18(i) that for any $n \in \mathbb{Z}$ the set $\mathbb{Z}_n$ is the set of equivalence classes in $\mathbb{Z}$ with respect to the equivalence relation $\equiv \pmod{n}$. In Examples A.2.16(i) and A.2.16(ii) we showed that the operations of modular addition and multiplication are well defined. In those examples we wrote $\llbracket k \rrbracket = \{\dots, k - n, k, k + n, \dots\}$ to denote the equivalence class of $k \pmod{n}$, and we wrote $\boxplus$ and $\boxtimes$ to denote modular addition and multiplication, respectively. It is common, however, for most mathematicians to be sloppy with notation and just write k to indicate $\llbracket k \rrbracket$, to write $+$ for $\boxplus$, and to write $\cdot$ (or just put letters next to each other) for $\boxtimes$. We also usually use this convenient shorthand.

Unexample B.2.5. The set $\mathbb{Z}_4 = \{0, 1, 2, 3\}$ with modular addition and multiplication is not a field because the element 2 has no multiplicative inverse. To see this, assume by way of contradiction that some element a is the multiplicative inverse of 2. In this case we have

$$2 = (a \cdot 2) \cdot 2 = a \cdot (2 \cdot 2) = a \cdot 0 = 0,$$

a contradiction.

In fact, if there is a $d \not\equiv 0, 1 \pmod{n}$ which divides n , a very similar argument shows that $\mathbb{Z}_n$ is not a field, because $d \not\equiv 0 \pmod{0}$ has no multiplicative inverse.

It is straightforward to verify that $\mathbb{Z}_n$ satisfies all the properties of a field except the existence of multiplicative inverses. In the special case that $n = p$ is prime, we do have multiplicative inverses, and $\mathbb{Z}_p$ is a field.

Proposition B.2.6. *If $p \in \mathbb{Z}$ is prime, then $\mathbb{Z}_p$ with the operations of modular addition and multiplication is a field.*

Proof. As mentioned above, it is straightforward to verify that all the axioms of a field hold except multiplicative inverses. We now show that nonzero elements of $\mathbb{Z}_p$ all have multiplicative inverses. Any element $a \in \mathbb{Z}$ with $a \not\equiv 0 \pmod{p}$ is relatively prime to p , which means that $\gcd(a, p) = 1$. By Proposition 15.2.8, this means there exist $x, y \in \mathbb{Z}$ such that $ax + py = 1$. This implies that $ax \equiv 1 \pmod{p}$, and hence x is the multiplicative inverse to a in $\mathbb{Z}_p$. $\square$



Topics in Matrix Analysis

It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.

—Emil Artin

C.1 Matrix Algebra

C.1.1 Matrix Fundamentals

Definition C.1.1. An $m \times n$ matrix A over $\mathbb{F}$ is a rectangular array of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

with m rows and n columns, where each entry a_{ij} is an element of $\mathbb{F}$. We also write $A = [a_{ij}]$ to indicate that the (i, j) entry of A is a_{ij} . We denote the set of $m \times n$ matrices over $\mathbb{F}$ by $M_{m \times n}(\mathbb{F})$.

Matrices are important because they give us a nice way to write down linear transformations explicitly (see Section 2.4). Composition of linear transformations corresponds to matrix multiplication (see Section 2.5.1).

Definition C.1.2. Let A be an $m \times r$ matrix with entries $[a_{ij}]$, and let B be an $r \times n$ matrix with entries $[b_{ij}]$. The product AB is defined to be the $m \times n$ matrix C with entries $[c_{ij}]$, where

$$c_{ij} = \sum_{k=1}^r a_{ik}b_{kj}.$$

Note that you can also consider c_{ij} to be the inner product of the i th row of A and the j th column of B . Given a scalar $s \in \mathbb{F}$ and a matrix $A \in M_{m \times n}(\mathbb{F})$ with entries $[a_{ij}]$, we define scalar multiplication of A by s to be the matrix sA whose (i, j) entry is sa_{ij} .

Definition C.1.3. The transpose A^T of a matrix $A = [a_{ij}]$ is defined to be the matrix

$$A^T = [a_{ji}]. \quad (\text{C.1})$$

The Hermitian conjugate A^H of A is the complex conjugate of the transpose

$$A^H = \overline{A}^T = [\overline{a_{ji}}]. \quad (\text{C.2})$$

It is straightforward to verify that for any $A, B \in M_{m \times n}(\mathbb{F})$ we have

$$\begin{aligned} (A^T)^T &= A, & (A+B)^T &= A^T + B^T, \\ (A^H)^H &= A, & \text{and} & \quad (A+B)^H = A^H + B^H. \end{aligned}$$

The following multiplicative property of transpose and conjugates can be verified by a tedious computation that we will not give here.

Proposition C.1.4. For any $m \times r$ matrix A and any $r \times n$ matrix B we have

$$(AB)^T = B^T A^T \quad \text{and} \quad (AB)^H = B^H A^H.$$

C.1.2 Matrix Algebra

Theorem C.1.5. For scalars $a, b, c \in \mathbb{F}$ and matrices A, B, C of the appropriate shapes for the following operations to make sense, we have the following rules of matrix algebra:

- (i) Addition is commutative: $A + B = B + A$.
- (ii) Addition is associative: $(A + B) + C = A + (B + C)$.
- (iii) Multiplication is associative: $(AB)C = A(BC)$.
- (iv) Matrix multiplication is left distributive: $A(B + C) = AB + AC$.
- (v) Matrix multiplication is right distributive: $(A + B)C = AC + BC$.
- (vi) Scalar multiplication is associative: $(ab)A = a(bA)$.
- (vii) Scalar multiplication is distributive over matrix addition: $a(A + B) = aA + aB$.
- (viii) Matrix multiplication is distributive over scalar addition: $(a + b)A = aA + bA$.

Proof. Associativity of matrix multiplication is proved in Corollary 2.5.3. The proof of the remaining properties is a tedious but straightforward verification that can be found in many elementary linear algebra texts. $\square$

Nota Bene C.1.6. Matrix multiplication is not commutative. In fact, if A and B are not square, and AB is defined, then the BA is not defined. But even if A and B are square, we rarely have equality for the two products AB and BA .

Proposition C.1.7. *If $A \in M_{m \times n}(\mathbb{F})$ satisfies $A\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in \mathbb{F}^n$, then $A = 0$.*

Proof. If $\mathbf{x} = \mathbf{e}_i$ is the i th standard basis vector, then $A\mathbf{e}_i$ is the i th column of A . Since this is zero for every i , the entire matrix A must be zero. $\square$

C.1.3 Inverses

Definition C.1.8. *A matrix $A \in M_n$ is invertible (or nonsingular) if there exists $B \in M_n$ such that $AB = BA = I$.*

Proposition C.1.9. *Matrix inverses are unique.*

Proof. If B and C are both inverses of A , then

$$B = BI = B(AC) = (BA)C = IC = C. \quad \square$$

Proposition C.1.10. *If $A, B \in M_n(\mathbb{F})$ are both invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.*

Proof. We have

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = I$$

and

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = I. \quad \square$$

C.2 Block Matrices

Partitioning a matrix into submatrices or *blocks* is often helpful. For example, we may partition the matrix A as

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}.$$

Here the number of rows in B and C are equal, the number of rows in D and E are equal, the number of columns in B and D are equal, and the number of columns in C and E are equal. There is no requirement that A , B , C , D , or E be square.

C.2.1 Block Matrix Multiplication

Suppose we are given two matrices A and B that are partitioned into blocks as below:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ A_{21} & A_{22} & \cdots & A_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mr} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{r1} & B_{r2} & \cdots & B_{rn} \end{bmatrix}.$$

If for every pair (A_{ik}, B_{kj}) the number of columns of A_{jk} equals the number of rows of B_{kj} , then the product of the matrices is formed in a manner similar to that of regular matrix multiplication. In fact, the (i, j) block of the matrix is equal to $\sum_{k=1}^r A_{ik} B_{kj}$.

Example C.2.1. Suppose

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Then the product is

$$AB = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Block matrix multiplication is especially useful when there are patterns (usually involving zeros or the identity in the matrices to be multiplied).

C.2.2 Block Matrix Identities

Lemma C.2.2 (Schur). *Let M be a square matrix with block form*

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}. \quad (\text{C.3})$$

If A , D , $A - BD^{-1}C$, and $D - CA^{-1}B$ are nonsingular, then we have two equivalent descriptions of the inverse of M , that is,

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \quad (\text{C.4})$$

and

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}. \quad (\text{C.5})$$

The matrices $A - BD^{-1}C$ and $D - CA^{-1}B$ are called the Schur complements of A and D , respectively.

Proof. These identities can be verified by multiplying (C.3), (C.4), and (C.5), respectively. $\square$

Lemma C.2.3 (Sherman–Morrison–Woodbury). *Let A and D be invertible matrices and B and D be given so that the sum $A + BD^{-1}C$ is nonsingular. If $D - CA^{-1}B$ is also nonsingular, then*

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Proof. This follows by equating the upper left blocks of (C.4) and (C.5). $\square$

C.3 Cross Products

The *cross product* of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ is defined to be the unique vector $\mathbf{x} \times \mathbf{y}$ with the property that $(\mathbf{x} \times \mathbf{y}) \cdot \mathbf{z} = \det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ for every $\mathbf{z} \in \mathbb{R}^3$. Here $\det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ means the determinant of the matrix formed by writing $\mathbf{x} = (x_1, x_2, x_3)^T$, $\mathbf{y} = (y_1, y_2, y_3)^T$, and $\mathbf{z} = (z_1, z_2, z_3)^T$ in terms of the standard basis and assembling these together to form

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix},$$

so we have

$$(\mathbf{x} \times \mathbf{y}) \cdot \mathbf{z} = \det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \det \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix},$$

where $\det(A)$ is the determinant of A ; see Sections 2.8 and 2.9. Geometrically, $|\det(\mathbf{x}, \mathbf{y}, \mathbf{z})|$ is the volume of the parallelepiped having $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as three of the sides; see Remark 8.7.6.

Proposition C.3.1. *For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^3$ and any $a, b \in \mathbb{R}$ the following properties hold:*

- (i) $\mathbf{x} \times \mathbf{y} = -\mathbf{y} \times \mathbf{x}$.
- (ii) *The cross product is bilinear; that is,*
 $(a\mathbf{x} + b\mathbf{z}) \times \mathbf{y} = a(\mathbf{x} \times \mathbf{y}) + b(\mathbf{z} \times \mathbf{y}) \quad \text{and} \quad \mathbf{x} \times (a\mathbf{y} + b\mathbf{z}) = a(\mathbf{x} \times \mathbf{y}) + b(\mathbf{x} \times \mathbf{z}).$
- (iii) $\mathbf{x} \times \mathbf{y} = \mathbf{0}$ if and only if $\mathbf{x} = c\mathbf{y}$ for some $c \in \mathbb{R}$.
- (iv) $\langle \mathbf{x} \times \mathbf{y}, \mathbf{x} \rangle = 0 = \langle \mathbf{x} \times \mathbf{y}, \mathbf{y} \rangle$; thus, $\mathbf{x} \times \mathbf{y}$ is orthogonal to the plane spanned by $\mathbf{x}$ and $\mathbf{y}$.
- (v) $\|\mathbf{x} \times \mathbf{y}\|$ is the area of the parallelogram having $\mathbf{x}$ and $\mathbf{y}$ as two of the sides.
- (vi) *The cross product of $\mathbf{x} = (x_1, x_2, x_3)$ with $\mathbf{y} = (y_1, y_2, y_3)$ can be computed as*

$$\mathbf{x} \times \mathbf{y} = \det \begin{pmatrix} x_2 & x_3 \\ y_2 & y_3 \end{pmatrix} \mathbf{e}_1 - \det \begin{pmatrix} x_1 & x_3 \\ y_1 & y_3 \end{pmatrix} \mathbf{e}_2 - \det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} \mathbf{e}_3.$$

Proof. The proof is a straightforward but messy computation. $\square$

Replacing $\mathbf{x}$ and $\mathbf{y}$ by linear combinations of $\mathbf{x}$ and $\mathbf{y}$ will change the cross product by the determinant of the transformation.

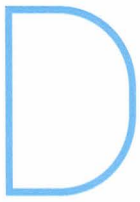
Proposition C.3.2. *Any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ define a linear transformation $B: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ by $\mathbf{e}_1 \mapsto \mathbf{x}$ and $\mathbf{e}_2 \mapsto \mathbf{y}$. Given any linear transformation $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$, the composition $B \circ \Phi$ defines another linear transformation $\mathbb{R}^2 \rightarrow \mathbb{R}^3$. Define $\mathbf{x}' = B \circ \Phi(\mathbf{e}_1)$ and $\mathbf{y}' = B \circ \Phi(\mathbf{e}_2)$. We have*

$$\mathbf{x}' \times \mathbf{y}' = \det(\Phi)(\mathbf{x} \times \mathbf{y}). \quad (\text{C.6})$$

Proof. Since we may factor Φ into elementary matrices, it suffices to show that (C.6) holds when Φ is elementary (see Section 2.7.1). If Φ is type I, then the determinant is -1 , and we have $\mathbf{x}' = \mathbf{y}$ and $\mathbf{y}' = \mathbf{x}$, so the desired result follows from (i) in Proposition C.3.1. If Φ is type II, corresponding to multiplication of the first row by α , then $\det(\Phi) = \alpha$, and it is easy to see that $\mathbf{x}' \times \mathbf{y}' = \alpha(\mathbf{x} \times \mathbf{y})$. Finally, consider the case where Φ is type III, corresponding to adding a scalar multiple of one row to the other row. Let us assume first that $\Phi = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$. We have $\mathbf{x}' = \mathbf{x} + \alpha\mathbf{y}$ and $\mathbf{y}' = \mathbf{y}$. We have $\det(\Phi) = 1$ and

$$\mathbf{x}' \times \mathbf{y}' = (\mathbf{x} + \alpha\mathbf{y}) \times \mathbf{y} = \mathbf{x} \times \mathbf{y} + \alpha\mathbf{y} \times \mathbf{y} = \mathbf{x} \times \mathbf{y}.$$

The case of $\Phi = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix}$ is essentially identical. $\square$



The Greek Alphabet

I fear the Greeks even when they bring gifts.
—Virgil

CAPITAL	LOWER	VARIANT	NAME
<i>A</i>	α		Alpha
<i>B</i>	β		Beta
Γ	γ		Gamma
Δ	δ		Delta
<i>E</i>	ϵ	ε	Epsilon
<i>Z</i>	ζ		Zeta
<i>H</i>	η		Eta
Θ	θ	ϑ	Theta
<i>I</i>	ι		Iota
<i>K</i>	κ	$\varkappa$	Kappa
Λ	λ		Lambda
<i>M</i>	μ		Mu
<i>N</i>	ν		Nu
Ξ	ξ		Xi
<i>O</i>	o		Omicron
Π	π	ϖ	Pi
<i>P</i>	ρ	ϱ	Rho
Σ	σ	ς	Sigma
<i>T</i>	τ		Tau
Υ	υ		Upsilon
Φ	ϕ	φ	Phi
<i>X</i>	χ		Chi
Ψ	ψ		Psi
Ω	ω		Omega

Bibliography

- [AB66] Edgar Asplund and Lutz Bungart. *A First Course in Integration*. Holt, Rinehart and Winston, New York, Toronto, London, 1966. [360]
- [Abb15] Stephen Abbott. *Understanding Analysis*. Undergraduate Texts in Mathematics. Springer, New York, second edition, 2015. [xiv, xvii]
- [Ahl78] Lars V. Ahlfors. *Complex Analysis. An Introduction to the Theory of Analytic Functions of One Complex Variable*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, third edition, 1978. [456]
- [Alu09] Paolo Aluffi. *Algebra: Chapter 0*. Volume 104 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2009. [624]
- [Ans06] Richard Anstee. The Newton-Raphson method. <https://www.math.ubc.ca/~anstee/math104/104newtonmethod.pdf>, 2006. Last accessed 5 April 2017. [315]
- [Art91] Michael Artin. *Algebra*. Prentice-Hall, Englewood Cliffs, NJ, 1991. [624]
- [Axl15] Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer, Cham, third edition, 2015. [85]
- [Ber79] S. K. Berberian. Classroom Notes: Regulated functions: Bourbaki's alternative to the Riemann integral. *Amer. Math. Monthly*, 86(3):208–211, 1979. [239]
- [BL06] Kurt Bryan and Tanya Leise. The \$25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Rev.*, 48(3):569–581, 2006. [517]
- [BM14] E. Blackstone and P. Mikusinski. The Daniell integral. *ArXiv e-prints*, January 2014. [360]
- [Bou76] N. Bourbaki. *Éléments de mathématique. Fonctions d'une variable réelle, Théorie élémentaire*. Hermann, Paris, 1976. [228]
- [Bre08] David M. Bressoud. *A Radical Approach to Lebesgue's Theory of Integration*. MAA Textbooks. Cambridge University Press, Cambridge, 2008. [360]

- [Bro00] Andrew Browder. Topology in the complex plane. *Amer. Math. Monthly*, 107(5):393–401, 2000. [406]
- [BT04] Jean-Paul Berrut and Lloyd N. Trefethen. Barycentric Lagrange interpolation. *SIAM Rev.*, 46(3):501–517, 2004. [611, 612]
- [CB09] Ruel V. Churchill and James Ward Brown. *Complex Variables and Applications*. McGraw-Hill, New York, eighth edition, 2009. [456]
- [CF13] Robert M. Corless and Nicolas Fillion. *A Graduate Introduction to Numerical Methods. From the Viewpoint of Backward Error Analysis*. With a foreword by John Butcher. Springer, New York, 2013. [551]
- [Cha95] Soo Bong Chae. *Lebesgue Integration*. Universitext. Springer-Verlag, New York, second edition, 1995. [332, 345, 360, 380]
- [Cha12] Francoise Chatelin. *Eigenvalues of Matrices*. Volume 71 of Classics in Applied Mathematics. With exercises by Mario Ahués and the author. Translated with additional material by Walter Ledermann; revised reprint of the 1993 edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012. [516]
- [Cha15] Tim Chartier. *When Life is Linear. From Computer Graphics to Bracketology*. Volume 45 of Anneli Lax New Mathematical Library. Mathematical Association of America, Washington, DC, 2015. [30]
- [Chi06] Carmen Chicone. *Ordinary Differential Equations with Applications*. Volume 34 of Texts in Applied Mathematics. Springer, New York, second edition, 2006. [315]
- [CM09] Stephen L. Campbell and Carl D. Meyer. *Generalized Inverses of Linear Transformations*. Volume 56 of Classics in Applied Mathematics. Reprint of the 1991 edition; corrected reprint of the 1979 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009. [517]
- [Con90] John B. Conway. *A Course in Functional Analysis*. Volume 96 of Graduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1990. [137, 176]
- [Con16] Keith Conrad. Differentiation under the integral sign. <http://www.math.uconn.edu/~kconrad/blurbs/analysis/diffunderint.pdf>, 2016. Last accessed 5 April 2017. [360]
- [Cos12] Christopher M. Cosgrove. Contour integration and Cauchy’s theorem, 2012. Last accessed 2 Jan 2017. [456]
- [Dem82] James Demmel. The condition number of similarities that diagonalize matrices. Technical report, Electronics Research Laboratory Memorandum, University of California at Berkeley, Berkeley, CA, 1982. [572]

- [Dem83] James Demmel. The condition number of equivalence transformations that block diagonalize matrix pencils. *SIAM J. Numer. Anal.*, 20(3):599–610, 1983. [572]
- [Dem97] James W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. [315]
- [Der72] William R. Derrick. *Introductory complex analysis and applications*. Academic Press, New York, 1972. [456]
- [Die60] J. Dieudonné. *Foundations of Modern Analysis*. Volume 10 of Pure and Applied Mathematics. Academic Press, New York, London, 1960. [228]
- [DR50] A. Dvoretzky and C. A. Rogers. Absolute and unconditional convergence in normed linear spaces. *Proc. Nat. Acad. Sci. U. S. A.*, 36:192–197, 1950. [213]
- [Dri04] Bruce K. Driver. Analysis tools with examples. <http://www.math.ucsd.edu/~bdriver/DRIVER/Book/anal.pdf>, 2004. Last accessed 5 April 2017. [380]
- [DS10] Doug Smith, Maurice Eggen, and Richard St. Andre. *A Transition to Advanced Mathematics*. Brooks Cole, seventh edition, 2010. [627]
- [Fit06] P. Fitzpatrick. *Advanced Calculus*. Pure and Applied Undergraduate Texts. American Mathematical Society, Providence, RI, 2006. [360]
- [FLH85] Richard Phillips Feynman, Ralph Leighton, and Edward Hutchings. *Surely You're Joking, Mr. Feynman! Adventures of a Curious Character*. WW Norton & Company, 1985. [289]
- [GC08] Ping Zhang Gary Chartrand, Albert D. Polimeni. *Mathematical Proofs: A Transition to Advanced Mathematics*. Pearson/Addison Wesley, second edition, 2008. [627]
- [Gre97] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*. Volume 17 of Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. [524]
- [GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. [315, 551]
- [Hal74] Paul R. Halmos. *Naive Set Theory*. Undergraduate Texts in Mathematics. Reprint of the 1960 edition. Springer-Verlag, New York, 1974. [627, 629, 648, 650]
- [Hal07a] Thomas C. Hales. The Jordan curve theorem, formally and informally. *Amer. Math. Monthly*, 114(10):882–894, 2007. [406]
- [Hal07b] Thomas C. Hales. Jordan's proof of the Jordan curve theorem. *Studies in Logic, Grammar, and Rhetoric*, 10(23):45–60, 2007. [406]

- [Her96] I. N. Herstein. *Abstract Algebra*. Prentice–Hall, Upper Saddle River, NJ, third edition, 1996. [624]
- [HH99] John Hamal Hubbard and Barbara Burke Hubbard. *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Prentice–Hall, Upper Saddle River, NJ, 1999. [406]
- [HRW12] Jeffrey Humpherys, Preston Redd, and Jeremy West. A fresh look at the Kalman filter. *SIAM Rev.*, 54(4):801–823, 2012. [286]
- [HS75] Edwin Hewitt and Karl Stromberg. *Real and Abstract Analysis. A Modern Treatment of the Theory of Functions of a Real Variable*. Volume 25 of Graduate Texts in Mathematics, Springer-Verlag, New York-Heidelberg, 1975. [201]
- [IM98] Ilse C. F. Ipsen and Carl D. Meyer. The idea behind Krylov methods. *Amer. Math. Monthly*, 105(10):889–899, 1998. [551]
- [Ise09] Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2009. [524]
- [JL97] Erxiong Jiang and Peter C. B. Lam. An upper bound for the spectral condition number of a diagonalizable matrix. *Linear Algebra Appl.*, 262:165–178, 1997. [572]
- [Jon93] Frank Jones. *Lebesgue Integration on Euclidean Space*. Jones and Bartlett Publishers, Boston, MA, 1993. [360]
- [JT98] G. F. Jónsson and L. N. Trefethen. A numerical analyst looks at the “cutoff phenomenon” in card shuffling and other Markov chains. In *Numerical analysis 1997 (Dundee)*, volume 380 of Pitman Res. Notes Math. Ser., pages 150–178. Longman, Harlow, 1998. [564]
- [KA82] L. V. Kantorovich and G. P. Akilov. *Functional Analysis*. Translated from the Russian by Howard L. Silcock. Pergamon Press, Oxford, Elmsford, NY, second edition, 1982. [315]
- [Kan52] L. V. Kantorovich. Functional analysis and applied mathematics. NBS Report 1509. U. S. Department of Commerce, National Bureau of Standards, Los Angeles, CA, 1952. Translated by C. D. Benster. [315]
- [Kan88] Shen Kangsheng. Historical development of the Chinese Remainder Theorem. *Archive for History of Exact Sciences*, 38(4):285–305, 1988. [606]
- [Kat95] Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995. [516]
- [Lay02] D.C. Lay. *Linear Algebra and Its Applications*. Pearson Education, 2002. [30, 85]

- [Leo80] Steven J. Leon. *Linear Algebra with Applications*. Macmillan, Inc., New York; Collier-Macmillan Publishers, London, 1980. [30, 85]
- [Lib73] Ulrich Libbrecht. *Chinese Mathematics in the Thirteenth Century*. Volume 1 of MIT East Asian Science Series. The Shu-shu chiu-chang of Ch'in Chiu-shao. M.I.T. Press, Cambridge, MA, London, 1973. [624]
- [LM06] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, Princeton, NJ, 2006. [517]
- [Mac00] C. R. MacCluer. The many proofs and applications of Perron's theorem. *SIAM Rev.*, 42(3):487–498, 2000. [517]
- [Mae84] Ryuji Maehara. The Jordan curve theorem via the Brouwer fixed point theorem. *Amer. Math. Monthly*, 91(10):641–643, 1984. [406]
- [MH99] Jerrold E. Marsden and Michael J. Hoffman. *Basic Complex Analysis*. W. H. Freeman and Company, New York, third edition, 1999. [456]
- [Mik14] P. Mikusinski. Integrals with values in Banach spaces and locally convex spaces. *ArXiv e-prints*, March 2014. [360]
- [Mor17] Sidney A. Morris. Topology without tears. <http://www.topologywithouttears.net/topbook.pdf>, 2017. Last accessed 5 April 2017. [239]
- [Mun75] James R. Munkres. *Topology: A First Course*. Prentice–Hall, Englewood Cliffs, NJ, 1975. [227, 239]
- [Mun91] James R. Munkres. *Analysis on Manifolds*. Advanced Book Program. Addison-Wesley Publishing Company, Redwood City, CA, 1991. [380]
- [NJN98] Gail Nord, David Jabon, and John Nord. The global positioning system and the implicit function theorem. *SIAM Rev.*, 40(3):692–696, 1998. [301]
- [Ort68] James M. Ortega. The Newton-Kantorovich theorem. *Amer. Math. Monthly*, 75:658–660, 1968. [315]
- [OS06] Peter J. Olver and Chehrzad Shakiban. *Applied Linear Algebra*. Pearson Prentice–Hall, Upper Saddle River, NJ, 2006. [30, 85]
- [Pal07] Richard S. Palais. A simple proof of the Banach contraction principle. *J. Fixed Point Theory Appl.*, 2(2):221–223, 2007. [315]
- [Pro08] S. David Promislow. *A First Course in Functional Analysis*. Pure and Applied Mathematics (Hoboken). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008. [131, 137, 176]
- [Rey06] Luc Rey-Bellet. Math 623 homework 3. http://people.math.umass.edu/~lr7q/ps_files/teaching/math623/PS3.pdf, 2006. Last accessed 5 April 2017. [360]

- [RF10] H. L. Royden and P. M. Fitzpatrick. *Real Analysis*. Prentice-Hall, New York, fourth edition, 2010. [335, 360]
- [Roy63] H. L. Royden. *Real Analysis*. The Macmillan Co., New York; Collier-Macmillan Ltd., London, first edition, 1963. [360]
- [RT90] Satish C. Reddy and Lloyd N. Trefethen. Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues. *Spectral and high order methods for partial differential equations* (Como, 1989). *Comput. Methods Appl. Mech. Engrg.*, 80(1-3):147–164, 1990. [565]
- [RT92] Satish C. Reddy and Lloyd N. Trefethen. Stability of the method of lines. *Numer. Math.*, 62(2):235–267, 1992. [565]
- [Rud87] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, third edition, 1987. [360]
- [Rud91] Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, second edition, 1991. [137, 176]
- [Sch12] Konrad Schmüdgen. *Unbounded Self-adjoint Operators on Hilbert Space*. Volume 265 of Graduate Texts in Mathematics. Springer, Dordrecht, 2012. [470]
- [Soh14] Houshang H. Sohrab. *Basic Real Analysis*. Birkhäuser/Springer, New York, second edition, 2014. [360]
- [Spi65] Michael Spivak. *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*. W. A. Benjamin, Inc., New York, Amsterdam, 1965. [406]
- [SS02] E. B. Saff and A. D. Snider. *Fundamentals of Complex Analysis with Applications for Engineering and Science*. Prentice-Hall, Pearson Education Inc., third edition, 2002. [456]
- [Ste98] G. W. Stewart. *Matrix Algorithms. Vol. I. Basic Decompositions*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998. [167]
- [Ste09] William Stein. *Elementary Number Theory: Primes, Congruences, and Secrets. A Computational Approach*. Undergraduate Texts in Mathematics. Springer, New York, 2009. [624]
- [Str80] Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York, London, second edition, 1980. [30, 85]
- [Str93] Gilbert Strang. The fundamental theorem of linear algebra. *Amer. Math. Monthly*, 100(9):848–855, 1993. [163]
- [Str10] Gilbert Strang. Linear algebra. Spring 2010. <http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/>, 2010. Last accessed 5 April 2017. [30, 85]

- [Tap09] Richard Tapia. Keynote lecture 4: If it's fast it must be Newton's method. In *Proceedings of the 15th American Conference on Applied Mathematics*, AMATH'09, pages 14–14, Stevens Point, WI, 2009. World Scientific and Engineering Academy and Society (WSEAS). [286]
- [Tay11] Joseph L. Taylor. *Complex Variables*. Volume 16 of Pure and Applied Undergraduate Texts. American Mathematical Society, Providence, RI, 2011. [456]
- [TB97] Lloyd N. Trefethen and David Bau, III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. [137, 315, 516, 547, 551]
- [TE05] Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra. The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, NJ, 2005. [564, 565, 572]
- [Tre92] L. N. Trefethen. Pseudospectra of matrices. In *Numerical Analysis 1991 (Dundee, 1991)*, volume 260 of Pitman Res. Notes Math. Ser., pages 234–266. Longman Sci. Tech., Harlow, 1992. [572]
- [Tre97] Lloyd N. Trefethen. Pseudospectra of linear operators. *SIAM Rev.*, 39(3):383–406, 1997. [572]
- [Van08] E. B. Van Vleck. On non-measurable sets of points, with an example. *Transactions of the American Mathematical Society*, 9(2):237–244, April 1908. [335]
- [Wik14] MathOverflow Community Wiki. Applications of the Chinese remainder theorem. <http://mathoverflow.net/questions/10014/applications-of-the-chinese-remainder-theorem>, 2009–2014. Last accessed 5 April 2017. [624]
- [Wil79] James H Wilkinson. Note on the practical significance of the Drazin inverse. Technical report, Stanford, CA, 1979. [517]
- [Wil07] Mark Wildon. A short proof of the existence of the Jordan normal form. <http://www.ma.rhul.ac.uk/~uvah099/Maths/JNFfinal.pdf>, 2007. Last accessed 5 April 2017. [517]
- [Wri02] Thomas G. Wright. Eigtool. <http://www.comlab.ox.ac.uk/pseudospectra/eigtool/>, 2002. Last accessed 5 April 2017. [572]
- [WT94] Elias Wegert and Lloyd N. Trefethen. From the Buffon needle problem to the Kreiss matrix theorem. *Amer. Math. Monthly*, 101(2):132–139, 1994. [572]

Index

- Abel–Weierstrass lemma, 413
- absolute condition number,
302
- addition
 - closure of, 5, 576
- adjoint, 121
- adjugate, 73, 76
- algebraic multiplicity, 143
- almost everywhere, 329, 332
 - convergence, 333
 - equal, 332
 - nonnegative, 340
- analytic function, 411, 414, 434
- angle preserving, 131
- arclength, 383
 - function, 384
 - parametrized by, 384
- argument principle, 445
- arithmetic-geometric mean inequality,
118
- Arnoldi
 - basis, 531
 - method, 547
- asymptotic behavior, 561
- automorphism
 - of a ring, 595
 - of a vector space, 38
- axiom of choice, 647

- Bézier curve, 54
- back substitution, 61
- Banach space, 211
- Banach-valued regulated integral
 - multivariable, 324
 - single-variable, 230
- barycentric
 - Lagrange interpolation,
610
 - weight, 610

- basic
 - columns, 65
 - variables, 63
- basis, 13
 - standard, 13, 51
- Bauer–Fike theorem, 560
- Beppo Levi theorem, 337
- Bernstein
 - operator, 82
 - polynomials, 54
- Bessel’s inequality, 97
- bijection, 38
- bijective, 38, 637
- bilinear, 89, 667
- block matrix
 - inverse, 666
 - multiplication, 665
- Bochner integral, 320
- Boolean ring, 575, 595
- boundary of a set, 192
- bounded
 - above, 645
 - functions, 113
 - linear transformation, 114, 186,
200, 235
 - linear transformation theorem, 213,
216
 - set, 197
- Brouwer fixed-point theorem, 225, 277

- calculus
 - fundamental theorem of, 403
- cancellation, 588
- canonical epimorphism, 41, 600
- Cantor
 - diagonal argument, 650
 - ternary set, 332
- cardinality, 648
- Cartesian product, 14, 19, 596, 630

- Cauchy
 - differentiation formula, 427
 - integral formula, 413, 424
 - sequence, 195, 196, 263
- Cauchy–Goursat theorem, 419, 421
- Cauchy–Riemann
 - equation, 410
 - equations, polar form, 451
 - theorem, 410
- Cauchy–Riemann equation, 410
- Cauchy–Schwarz inequality, 91
- Cayley–Hamilton theorem, 479
- chain, 20, 647
- chain rule
 - for holomorphic functions, 412
 - Fréchet derivative, 259
- change of basis, 53
- change of variables, 263, 351
- characteristic polynomial, 143
- Chinese remainder
 - problem, 602
 - theorem, 80, 601, 604
- classical adjoint, 76
- closed
 - n -interval, 321
 - ball, 191
 - set, 190
- closure
 - of a set, 192
 - of operations, 5, 576
- codomain, 635
- cofactor, 73
 - expansion, 65, 75
- column space, 126
- commutative
 - ring, 575
- commutative diagram, 638
- compact, 179, 203
 - support, 122
- complementary
 - subspace, 16
- complete, 198
- completion, 361
- complex
 - conjugate, 654
 - numbers, 653
- composition, 36, 636
 - of continuous functions, 187
- computer aided design, 58
- condition, 302
- condition number
 - of a matrix, 305
 - absolute, of a function, 302
 - eigenvalue problem, 307
 - of a matrix, 307
 - relative, of a function, 303
- congruent
 - modulo n , 631
 - modulo an ideal, 598
- connected, 179, 224
 - component, 396
- conservative vector field, 388
- continuity
 - uniform, 199
- continuous
 - at a point, 185
 - function, 185, 188
 - linear extension theorem, 213, 216
 - Lipschitz, 186, 200, 235, 254, 278
 - pointwise, 187
 - uniformly, 199
- continuously differentiable, 266
- contour, 416
 - integral, 416, 417
 - simple closed, 419
- contraction
 - mapping, 278
 - mapping principle, 279
 - mapping principle, uniform, 282
 - mapping, uniform, 282
- convergence
 - absolute, 212
 - almost everywhere, 333
 - linear, 286
 - of sums, 212
 - pointwise, 210
 - quadratic, 286
 - uniform, 210, 413
 - uniform on compact subsets, 263
- convex, 261
- coordinate
 - change of, 299
- coordinates
 - change of, 48, 351
 - hyperbolic, 360
 - hyperspherical, 356

- in a basis, 47
 - polar, 352
 - spherical, 354
- coset, 23, 598
 - operations, 24, 599
- countable, 648
- cover
 - open, 203
- Cramer's rule, 77, 217
- cross product, 667
- CRT, *see* Chinese remainder theorem
- curl, 402
- curve
 - differentiable, 242
 - fitting, 129
 - piecewise-smooth, 382
 - positively oriented, 398
 - simple closed, 381
 - smooth, 382
 - smooth parametrized, 381
 - smooth, oriented, 382
 - smooth, unoriented, 382
- cutoff phenomenon, 564
- Daniell integral, 320, 328
- data compression, 168
- De Morgan's Laws, 630
- decay matrix, 564
- decomposition
 - LU, 62
 - polar, 165
 - QR, 103
 - singular value, 162
 - Wedderburn, 500
- dense, 190
- derivative, 242
 - directional, 244
 - higher, 266
 - linearity, 256
 - of a complex function, 408
 - of a parametrized curve, 242
 - second, 266
- determinant, 65
- De Moivre's formula, 655
- diagonal matrix, 162
- diagonalizable, 151
 - orthonormally, 157
- diffeomorphism, 349
- differentiable
 - complex function, 408
 - continuously, 253
 - function, 241, 246, 252
- dimension, 18
 - formula, 46
- direct sum, 14, 16
- Dirichlet function, 332
- divides, 585
- division property, 583
- domain
 - Euclidean, 583
 - of a function, 635
- dominated convergence theorem, 342
- dot product, 89
- Drazin inverse, 500, 501
- dual space, 248
- Eckart–Young, Schmidt, Mirsky theorem, 167
- EEA, *see* extended Euclidean algorithm, 587
- eigenbasis, 151
- eigennilpotent, 483
- eigenprojection, 463, 475
- eigenspace, 140
 - generalized, 465, 468, 486
- eigenvalue, 140
 - semisimple, 492
 - simple, 307, 496
- eigenvector
 - generalized, 468
- elementary
 - matrix, 59
 - product, 68
- empty set, 627
- equivalence
 - modulo a subspace, 22
 - class, 22, 598, 632
 - modulo n , 21, 631, 633
 - modulo an ideal, 598
 - relation, 598, 631
- Euclidean
 - algorithm, 586
 - domain, 573, 583
 - extended algorithm, 587
- Euler's formula, 411, 415, 655
- extension by zero, 325

- extension theorem, 18, 29
- exterior of a simple closed curve, 397
- extreme value theorem, 205
- Fatou's lemma, 340
- field, 659
- finite intersection property, 206
- first isomorphism theorem
 - for rings, 600
 - for vector spaces, 42
- fixed point, 278
- fomula
 - Euler, 411
- formal power series, 576
- Fourier transform, discrete, 622
- Fréchet derivative, 246, 252
 - chain rule, 259
 - higher-order, 266
 - product rule, 257
 - real finite-dimensional case, 246
- Fredholm
 - alternative, 136
 - integral transform, 281
- free variables, 63
- Frobenius
 - inner product, 90, 136
- Fubini's theorem, 320, 344
 - for step functions, 372
- function, 635
 - analytic, 411, 413
 - entire, 408
 - holomorphic, 408
 - integrable (Daniell–Lebesgue), 330
 - regulated integrable, 324, 375
 - Riemann integrable, 320, 325, 340
 - smooth, 266
- fundamental
 - subspaces theorem, 124
 - theorem of algebra, 430, 591
 - theorem of arithmetic, 590
 - theorem of calculus, 262
 - theorem of calculus for line
 - integrals, 388
- fundamental theorem
 - of calculus, 403
- Gamma function, 359
- Garner's formula, 606
- Gauss' mean value theorem, 426
- Gauss–Seidel
 - convergence, 524
 - method, 522
- gcd, 586
- generalized
 - eigenspace, 465, 468, 486
 - Heine–Borel theorem, 208
 - inverse, *see* pseudoinverse
 - Leibniz integral rule, 348
 - minimal residual method (GMRES), 530
 - step function, 379
- geometric multiplicity, 141
- Givens rotation, 545
- GMRES, 530, 533
- Google, 498
- Gram–Schmidt, 99
 - modified, 101
- graph, 635
- greatest
 - common divisor, 586
 - element, 646
 - lower bound, 645
- Green's theorem, 381, 399
- Hölder's inequality, 118
- Heine–Borel theorem
 - generalized, 208
 - on $\mathbb{R}^n$, 204
- Hermitian conjugate, 84, 664
- Hessian matrix, 266
- Hilbert's identity, 471
- Hilbert–Schmidt
 - inner product, 90
- holomorphic
 - function, 408, 434
 - open mapping theorem, 445, 450
- homeomorphism, 222
- Householder, 105
- Householder transformation, 105, 106
- hyperbolic coordinates, 360
- hyperplane, 321
- hyperspherical coordinates, 356
- ideal, 578
 - generating set, 581
- identity map, 39

- ill conditioned, 303
- image, 594, 635
- implicit function theorem, 294
- index of a matrix, 466
- index set, 639
- indicator function, 228, 323
- induced
 - metric, 208
 - norm on linear transformations, 114
 - norm, from an inner product, 111
- induction, 644
- infimum, 645
- inherited metric, 208
- injective, 637
- inner product, 88
 - Frobenius, 90, 136
 - Hilbert–Schmidt, 90
 - positivity of, 91
 - space, 89
 - standard, 89
- integers, 628
- integrable
 - function, Daniell–Lebesgue, 330
 - function, regulated, 324
 - function, Riemann, 320, 325, 340
 - function, regulated, 375
 - on an unbounded domain, 337
- integral
 - Bochner, 320
 - contour, 417
 - Daniell, 320, 328
 - iterated, 345
 - Lebesgue, 320, 328
 - mean value theorem, 262
- interior
 - of a set, 182
 - of a simple closed curve, 397
- interior point, 182
- intermediate value theorem, 225
- intersection of sets, 630
- invariant subspace, 147, 462, 468
- inverse function theorem, 298
- invertible, 38
 - element, 578
 - left, 642
 - right, 642
- involution, 172
- irreducible
 - element, of a ring, 588
 - matrix, 497
- isolated point, 189
- isometric embedding, 362
- isometry, 116
- isomorphic rings, 596
- isomorphism
 - of rings, 595, 596
 - of vector spaces, 36, 38
- iterated integral, 345
- iterative
 - numerical methods, 519
 - solvers, 521
- iterative methods, 520
- Jacobi method, 522
 - convergence, 524
- Jacobian
 - determinant, 297
 - matrix, 249
- Jordan
 - canonical form, 140, 506
 - curve theorem, 396
- Jordan normal form, 480
- Kantorovich, 315
- kernel, 35, 594
- Kreiss constant, 562
- Kreiss matrix theorem, 563
- Kronecker delta, 95
- Krylov
 - basis, 527
 - methods, 526
 - solution, 529
 - subspace, 506, 527
- Lagrange
 - basis, 603
 - decomposition, 602
 - interpolant, 603
 - interpolant, barycentric form, 610
 - interpolation, 610
 - interpolation, barycentric, 610
- Lagrange–Hermite interpolation, 616, 617
- Lanczos method, 548
- Laplace equation, 145

- Laurent
 - expansion, 436
 - polynomial, 576
 - series, 433, 436
 - series, principal part, 438
- law of cosines, 92
- leading entry, 61
- least
 - element, 647
 - upper bound, 645
 - upper bound property, 646
- least squares, 127
- Lebesgue
 - integrable, 367
 - integral, 320, 328
 - number, 206
- left
 - eigenvector, 154
 - invertible, 642
 - pointing normal vector, 398
- Legendre polynomial, 101, 453
- Leibniz integral rule
 - generalized, 348
- Leibniz's integral rule, 346
- Leontief, Wassily, 520
- level set, 191, 293, 635
- limit
 - inferior, 340
 - of a function, 188
 - of a sequence, 193
 - point, 189
 - pointwise, 210
 - superior, 340
- line fitting, 129
- line integral, 386
 - of scalar field, 386
 - of vector field, 387
- linear
 - approximation, 242
 - combination, 10
 - convergence, 286
 - dependence, 12
 - extension, 230
 - functional, 121
 - homogeneous system, 40
 - independence, 12
 - operator, 33
 - ordering, 644
 - system, 32, 58, 105, 519, 520, 527, 530
 - homogeneous, 35, 40
 - overdetermined, 127
 - transformation, 31–33
 - bounded, 114
 - composition of, 51
 - matrix representation, 49
- Liouville's theorem, 430
- Lipschitz
 - continuous, 186, 200, 235, 254, 278
 - locally at a point, 254
- lower bound, 645
- l.u.b. property, *see* least upper bound property
- LU decomposition, 62
- manifold, 381
- map, 635
- Markov chain, 564
- matrix, 663
 - ∞ -norm, 116
 - 1-norm, 116
 - augmented, 60
 - block, 665
 - decay, 564
 - diagonalizable, 151
 - diagonally dominant, 524
 - Hermitian conjugate, 664
 - index of, 466
 - invertible, 665
 - irreducible, 497
 - minor, 73
 - nilpotent, 171
 - nonnegative, 494
 - nonsingular, 52, 665
 - normal, 158
 - orthogonal, 98
 - orthonormal, 98, 132
 - positive, 494
 - primitive, 497
 - product, 663
 - row equivalent, 61
 - semisimple, 151
 - similar, 53
 - simple, 151
 - sparse, 519

- transition, 53
- transpose, 664
- tridiagonal, 145
- unitary, *see* matrix, orthonormal
- maximal element, 646
- maximum modulus principle, 431
- mean value theorem, 260
 - Gauss's, 426
 - integral, 262
- measurable
 - function, 333
 - set, 331, 333
- measure, 323
 - zero, 331
 - zero, alternative definition, 364
- meromorphic, 439
- method of successive approximations, 280
- metric, 179, 180
 - Cartesian product, 181
 - discrete, 181
 - Euclidean, 180
 - French railway, 233
 - induced, 208
 - inherited, 208
 - normalized, 181
 - space, 180
- minimal polynomial, 526
- minimum modulus principle, 453
- Minkowski's inequality, 119
- Mirsky, Schmidt, Eckart–Young theorem, 167
- modulus
 - componentwise, of a matrix, 513
 - of a complex number, 654
- monic polynomial, 143, 586
- monotone
 - convergence theorem, 336
 - decreasing, 336
 - increasing, 201, 336
- Moore–Penrose pseudoinverse, 126, 166
- multiplication
 - closure of, 576
- multiplicity
 - of a zero, 435, 446
 - of an eigenvalue
 - algebraic, 143, 156, 469
 - geometric, 141, 156, 467
- natural numbers, 628
- neighborhood, 182
- Neumann series, 215
- Newton
 - decomposition, 604
 - interpolation, 611
- Newton's method
 - scalar version, 287
 - vector version, 291
- Newton–Kantorovich theorem, 292, 315
- noise canceling, 10
- nonsingular matrix, 665
- norm, 111
 - L^p -, 113
 - p -, 112
 - Euclidean, 111
 - Frobenius, 113, 168
 - induced
 - from an inner product, 111
 - on linear transformations, 114
 - Manhattan, 111
 - matrix, 115
 - operator, 114
 - sup, 113
 - taxicab, 111
- normal
 - distribution, 359
 - equation, 128
 - matrix, 158
- normed linear space, 111
- null space, 35
- nullity, 43
- numerical
 - linear algebra, 62
- oblique projection, 460
- one-to-one, 637
- onto, 637
- open
 - ball, 182
 - cover, 203
 - set, 182
- order
 - of a nilpotent operator, 485
 - of a pole of a function, 439
 - of a zero of a function, 435
- orderings, 643

- orthogonal, 91
 - complement, 105, 123
 - projection, 93, 96
- orthonormal, 87
 - matrix, 98, 132
 - set, 95
 - transformation, 97
- orthonormally
 - diagonalizable, 157, 158
 - similar, 155
- outer product expansion,
 - 164, 174
- overdetermined, 127
- PageRank algorithm, 498
- parallelogram identity, 131
- parametrized
 - contour, 416
 - curve
 - equivalent, 382
 - smooth, 381
 - manifold, 389
 - equivalent, 389
 - measure of, 394
 - oriented, 389
 - tangent space, 390
 - unoriented, 390
 - surface, 389
- parametrized manifold, 381
- partial
 - derivative, 245
 - k th-order, 266
 - in a Banach space,
 - 255
 - ordering, 646
 - sums, 212
- partially ordered set, 646
- partition, 22, 598, 633
- path, 416
 - connected, 226
 - independent, 420
- periodic function, 236
- permutation, 66
 - even, 67
 - inversion, 67
 - odd, 67
 - sign, 67
 - transposition, 67
- Perron
 - root, 496
 - theorem, 496
- Perron–Frobenius
 - eigenvalue, 496
 - theorem, 497
- piecewise smooth, 382, 416
- pigeonhole principle, 652
- polar decomposition, 165
- polarization identity, 131
- pole of a function, 439
- polygonal path, 421
- polynomial
 - monic, 143, 586
 - ring, 575
- poset, *see* partially ordered set
- positive
 - definite, 159
 - semidefinite, 159
- potential function, 388
- power
 - method, 492
 - set, 628
- preimage, 186, 635
- preimage of a function, 187
- prime, 588
- primitive
 - matrix, 497
- primitive root of unity, 657
- principal ideal domain
 - Euclidean domain is a, 585
- projection, 136, 460, 639
 - canonical, 597
 - complementary, 460
 - map, 33
 - spectral, 475
- pseudoeigenvalue, 554
- pseudoeigenvector, 555
- pseudoinverse
 - Drazin, 500, 501
 - Moore–Penrose, 166
- pseudospectral
 - radius, 561
- pseudospectrum, 554
- Pythagorean
 - law, 92
 - theorem, 97

QR

- decomposition, 102, 103
- decomposition, reduced, 103
- iteration, 538

quadratic convergence, 286

quasi-Newton method, 290

quotient

- of a ring by an ideal, 598
- of a vector space by a subspace, 23

R -linear combination, 581

radius

- pseudospectral, 561

radius of convergence, 414

range, 35

rank, 43

rank-nullity theorem, 43

Rayleigh quotient, 174

reduced row echelon form, 61

REF, *see* row echelon form

reflection, 172

regulated

- integrable functions, 324
- integral
 - multivariable, Banach valued, 324
 - single variable, Banach valued, 228

relation, 631

relative condition number, 303

relatively prime, 586

reparametrization, 382

replacement theorem, 17

residual vector, 93

residue

- of a function, 440
- theorem, 442

resolvent, 470

- of an operator, 470
- set, 141, 470

reverse Fatou lemma, 358

Riemann integral, 325

Riemann's theorem, 426

Riesz representation theorem, 120

right

- eigenvector, 154
- invertible, 642

ring, 574

- commutative, 575, 576
- homomorphism, 592
- of polynomials in a matrix, 576

Ritz eigenvalues, 547

root

- of unity, 657
- simple, 304

rotation map, 34

Rouché's theorem, 448

row

- echelon form, 61
- echelon form, reduced, 61
- operations, 58
- reduction, 58

RREF, *see* row echelon form, reduced

Russell's paradox, 629

scalar, 4

Schmidt, Mirsky, Eckart–Young theorem, 167, 168

Schur

- complements, 666
- form, 140, 156
- lemma, 155

second

- barycentric form, 611
- derivative, 266
- isomorphism theorem, 45

segment, 648

self-adjoint, 155

seminorm, 111, 361

semisimple

- eigenvalue, 492
- matrix, 151
- spectral mapping theorem, 153

separated

- spectrally, 540

sequence, 193

- Cauchy, 195, 263
- convergent, 193
- uniformly convergent, 210

sequentially compact, 206

series, 212

sesquilinear, 89

set, 627

- complement, 630
- difference, 630

- intersection, 630
 - of measure zero, 331
 - partition, 633
 - union, 629
- Sherman–Morrison–Woodbury lemma, 666
- shifting
 - QR iteration, 550
- sign of a complex number, 656
- similar matrix, 53
- simple
 - closed curve, 381
 - eigenvalue, 307, 496
 - matrix, 151
 - pole, 439
 - region, 398
 - root, 304
- simply connected, 397
- singular
 - matrix, 52
 - value decomposition, 162
 - compact form, 164
 - values, 162
- singularity
 - essential, 439
 - isolated, 438
 - removable, 439
- skew-Hermitian, 171
- small gain theorem, 171
- smooth, 266
- SOR, *see* successive overrelaxation
- span, 10
- spectral
 - condition number, 560
 - decomposition theorem, 487
 - mapping theorem, 489
 - radius, 474, 478
 - resolution formula, 478
 - theorem
 - Hermitian matrices, 157
 - normal matrices, 158
- spectrally separated, 540
- spectrum, 141, 470
- spherical coordinates, 354
- Spijker's lemma, 566
- Stein map, 34
- step function, 228, 323
 - generalized, 379
 - integral of, 229, 324
- Stokes' theorem, 402
- subcover
 - finite, 203
- subdivision
 - generalized, 321
 - of intervals, 321
 - one-dimensional, 228
- submatrix, 73
- submultiplicative property, 115
- subsequence, 197
- subset, 628
 - proper, 628
- subspace, 7
 - complementary, 14, 16
 - invariant, 147
 - proper, 9
 - trivial, 9
- substitution formula, 349
- successive approximation, 280
- successive overrelaxation
 - convergence, 525
- superposition, 9
- support, 635
- supremum, 645
- surface
 - parametrized, 389
- surjective, 637
- SVD, *see* singular value decomposition
- Sylvester map, 34
- symmetric group, 66
- tangent
 - to a curve, 382
 - unit, 382
- tangent vector, 243
- Taylor
 - formula, 268, 269
 - remainder, 271
 - series for holomorphic functions, 433
- topological
 - properties, 179, 222, 223
 - spaces, 186

- topologically equivalent, 219, 220
 - norms, 134, 221
- topology, 179, 186
 - induced, 219
- total
 - derivative, 246
 - ordering, 643
- totally bounded, 206, 208
- transient behavior, 561
- transition matrix, 48, 53
- transpose, 664
- transposition matrix, 59
- tridiagonal matrix, 145
- unbounded linear transformation, 235
- uncountable, 648
- uniform
 - continuity, 195, 199
 - contraction mapping, 282
 - contraction mapping principle, 282
 - convergence, 210, 413
 - convergence on compact subsets, 263
 - properties, 223
- union of sets, 629
- unit, 578
 - normal, 391, 392
 - tangent, 382
 - vector, 91
- unitarily similar, 155
- unitary matrix, *see* orthonormal matrix
- unity
 - ring, 577
- universal property of the product, 639
- upper bound, 645
- valuation, 583
- variables
 - change of, 351
- vector, 4
 - field, 387
 - conservative, 388
 - length of, 91
 - space, 4
 - unit, 91
- vectors
 - angle between, 92
- volume, 331, 351
 - n -dimensional, 323
 - of n -ball, 360
 - of a parallelepiped, 394
- Wedderburn decomposition, 500
- well conditioned, 303
- well-defined functions, 640
- well-ordering axiom, 644
- well-ordering principle, 647
- winding number, 441
- WOA, *see* well-ordering axiom
- Wronskian, 85
- Young's inequality, 118
- zero
 - and pole counting formula, 446
 - divisors, 583
 - of order n , 435
- Zorn's lemma, 20, 647

